PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Deep learning approach for automatic assessment of schizophrenia and bipolar disorder in patients using R-R intervals

Kamil Książek[1]*, **Wilhelm Masarczyk**[2], **Przemysław Głomb**[3], **Michał Romaszewski**[3], **Krisztián Buza**[4,5], **Przemysław Sekuła**[3], **Michał Cholewa**[3], **Katarzyna Kołodziej**[3], **Piotr Gorczyca**[2], **Magdalena Piegza**[2]

1 Faculty of Mathematics and Computer Science, Jagiellonian University, Kraków, Poland, **2** Department of Psychiatry, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Tarnowskie Góry, Poland, **3** Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Gliwice, Poland, **4** Faculty of Finance and Accountancy, Budapest University of Economics and Business, Budapest, Hungary, **5** BioIntelligence Group, Department of Mathematics-Informatics, Sapientia Hungarian University of Transylvania, Targu Mures, Romania

* kamil.ksiazek@uj.edu.pl

OPEN ACCESS

Citation: Książek K, Masarczyk W, Głomb P, Romaszewski M, Buza K, Sekuła P, et al. (2025) Deep learning approach for automatic assessment of schizophrenia and bipolar disorder in patients using R-R intervals. PLoS Comput Biol 21(9): e1012983. https://doi.org/10.1371/journal.pcbi.1012983

Editor: Daniele Enrico Schiavazzi, University of Notre Dame, UNITED STATES OF AMERICA

Received: March 20, 2025

Accepted: August 13, 2025

Published: September 3, 2025

Data availability statement: All data used in experiments are available on Zenodo at link https://doi.org/10.5281/zenodo.8171266 while the source code for selected experiments is publicly available on a GitHub repository at https://github.com/iitis/Polar-HRV-classification.

Funding: The study of KK was partially supported by the flagship project entitled

Abstract

Schizophrenia and bipolar disorder are severe mental illnesses that significantly impact quality of life. These disorders are associated with autonomic nervous system dysfunction, which can be assessed through heart activity analysis. Heart rate variability (HRV) has shown promise as a potential biomarker for diagnostic support and early screening of those conditions. This study aims to develop and evaluate an automated classification method for schizophrenia and bipolar disorder using short-duration electrocardiogram (ECG) signals recorded with a low-cost wearable device. We conducted classification experiments using machine learning techniques to analyze R-R interval windows extracted from short ECG recordings. The study included 60 participants—30 individuals diagnosed with schizophrenia or bipolar disorder and 30 control subjects. We evaluated multiple machine learning models, including Support Vector Machines, XGBoost, multilayer perceptrons, Gated Recurrent Units, and ensemble methods. Two time window lengths (about 1 and 5 minutes) were evaluated. Performance was assessed using 5-fold cross-validation and leave-one-out cross-validation, with hyperparameter optimization and patient-level classification based on individual window decisions. Our method achieved classification accuracy of 83% for the 5-fold cross-validation and 80% for the leave-one-out scenario. Despite the complexity of our scenario, which mirrors real-world clinical settings, the proposed approach yielded performance comparable to advanced diagnostic methods reported in the literature. The results highlight the potential of short-duration HRV analysis as a cost-effective and accessible tool for aiding in the diagnosis of schizophrenia and bipolar disorder. Our findings support the feasibility of using wearable ECG devices and machine learning-based classification for psychiatric screening, paving the way for further research and clinical applications.

"Artificial Intelligence Computing Center Core Facility" from the DigiWorld Priority Research Area within the Excellence Initiative – Research University program at Jagiellonian University in Krakow. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Mental health conditions such as schizophrenia and bipolar disorder can be challenging to diagnose, but early detection is crucial to better outcomes. In our study, we explored a simple but powerful idea: using heart rate variability (HRV), the natural fluctuations in time between heartbeats, as a potential indicator of these disorders. As HRV reflects the balance of the autonomic nervous system, research has shown that people with schizophrenia and bipolar disorder often have distinct HRV patterns, suggesting that heart activity analysis could support mental health screening. We tested whether a wearable heart monitor, similar to a fitness tracker, could help recognize these patterns using artificial intelligence. By analyzing heart activity records from 60 individuals, some with these conditions and others without, we trained several machine learning models to differentiate between them. Our approach successfully classified individuals with about 80% accuracy. Although not a replacement for clinical diagnosis, heart monitoring combined with artificial intelligence could help detect signs of mental illness, leading to faster interventions. Our work opens the door for further studies and, ultimately, the development of easy-to-use tools that could support doctors and improve patient care.

1. Introduction

Schizophrenia and bipolar disorder (BD) are both severe, chronic mental disorders with a high population prevalence rate of more than 1% [1] and a well-documented increasing tendency in the case of schizophrenia [2]. Both diseases have a significant negative impact on patients' quality of life [2], including social interactions, employment levels and income [3]. Both diseases overlap in symptoms and pathophysiology and can be discussed in terms of spectrum and continuity [4]. Diagnosis of the disease in its early, prodromal stage, when treatment is most effective, is challenging [5]. Diagnosis usually occurs only when symptoms become more pronounced and aggravation of symptoms in these two diseases is one of the main reasons for psychiatric ward admission [6]. This increases treatment costs and the burden placed on the mental healthcare system [7]. Thus, early diagnosis and intervention are crucial for the improvement of treatment costs and outcomes [8]. Current diagnostic methods are based purely on clinical interviews by trained physicians [9,10]. It is time- and resource-consuming and open to debate whether this approach is truly objective [11,12].

The clinical diagnosis may change over time, with a retrospective consistency (a concept similar to sensitivity) of 91.5% and 89.3% in a clinical setting for schizophrenia and BD, respectively. In addition, the consistency of the diagnosis is significantly different when the patients are assessed in psychiatric emergency and out patient settings, with a retrospective consistency lower than 70% [13]. Although the results in a clinical setting could be considered satisfactory and provide a workable baseline, there is a desire to improve accuracy and achieve even better outcomes. More importantly, the findings outside the clinical setting highlight the importance of improving the overall quality of diagnostics, regardless of location, using accessible and user-friendly on-site diagnostic tools.

Various sources of biological data are taken into consideration to enhance and objectify the psychiatric diagnostic process. Among them, broad-ranging approaches, including Magnetic Resonance Imaging (MRI) [14,15], Electroencephalography (EEG) [16,17], functional MRI (fMRI) [18,19], genetic analysis [20], and Autonomic Nervous System (ANS)-mediated physiological patterns such as R-R interval derived heart rate variability (HRV) [21] offer a rich source of data on the functioning and structure of the brain. This complexity and richness of data pose a challenge in itself, e.g. in cases such as MRI and genetic profiling, as these types of data have a multidimensional character [22,23]. This can lead to a significantly larger number of false associations due to a higher degree of confounding effects [24] and makes it challenging and resource-intensive to identify the desired features [25,26]. Most importantly, global economic reality limits the access of most people to clinical settings with expensive equipment and trained personnel and consequently, to effective psychiatric care [27,28].

A possible solution to this problem might be the use of a diagnostic approach based on physiological biomarkers such as physical activity [29], galvanic skin response [30], pupillary response [31] or HRV [32]. These biomarkers can be easily obtained through currently available wearable devices [33–35], fairly inexpensive technology that can be used even outside the clinical setting [36]. Among these biomarkers, HRV stands out as particularly promising. Both schizophrenia and BD share numerous common genetic, endophenotypic, and phenomenological traits [37,38]. One such particular trait is an ANS dysregulation that can be observed through changes in HRV [21,32]. The classification of HRV patterns in a given individual could provide an objective source of symptom information within the diagnostic process [32,39,40]. A robust wearable device-based classification could significantly support the low-cost, widespread early diagnosis of schizophrenia and BD.

HRV-based wearable diagnostic systems have received significant attention in recent years. The authors of [33] propose a Mobile Health (mHealth) wearable device, which collects HRV, electrodermal activity (EDA) and movement data. After six days of wearing, the recorded data showed a statistically significant difference in HRV values between participants with schizophrenia and healthy controls, confirming previous results. However, this analysis was limited to long recording windows (> 8h) and did not consider the classification accuracy in a diagnostic setting. A similar approach is used by [41], where a disposable adhesive patch HR and activity sensor worn for several days are used to classify whether a patient is diagnosed with schizophrenia or belongs to a healthy control. This approach has achieved a promising AUC (area under a receiver operating characteristic curve) score of 0.96 for eight-day windows and 0.91 for two-day windows. However, this method is possibly limited by the manual choice of features, the lack of hyperparameter optimization for the classifier, and is restricted to a minimum of two days of patient observation.

A different approach is used by [42], which uses short time frame ECG measurements (25-minute windows), together with structured yoga exercises. The classification, based on the thresholding Z-score computed from Linear Discriminant Analysis (LDA), achieves a promising score of sensitivity of 91% and specificity of 81%. However, the sample size is limited (only 12 participants with schizophrenia) and, possibly because of that, no cross-validation protocol is mentioned, which severely limits the generalization of the reported scores. A wearable HRV-based setup was also successfully tested for hallucination spectrum experiences (HSE) and paranoia assessment [43], for the detection of stress in working people [44], as well as for the monitoring of symptoms in early psychosis [45] and anxiety disorders [46].

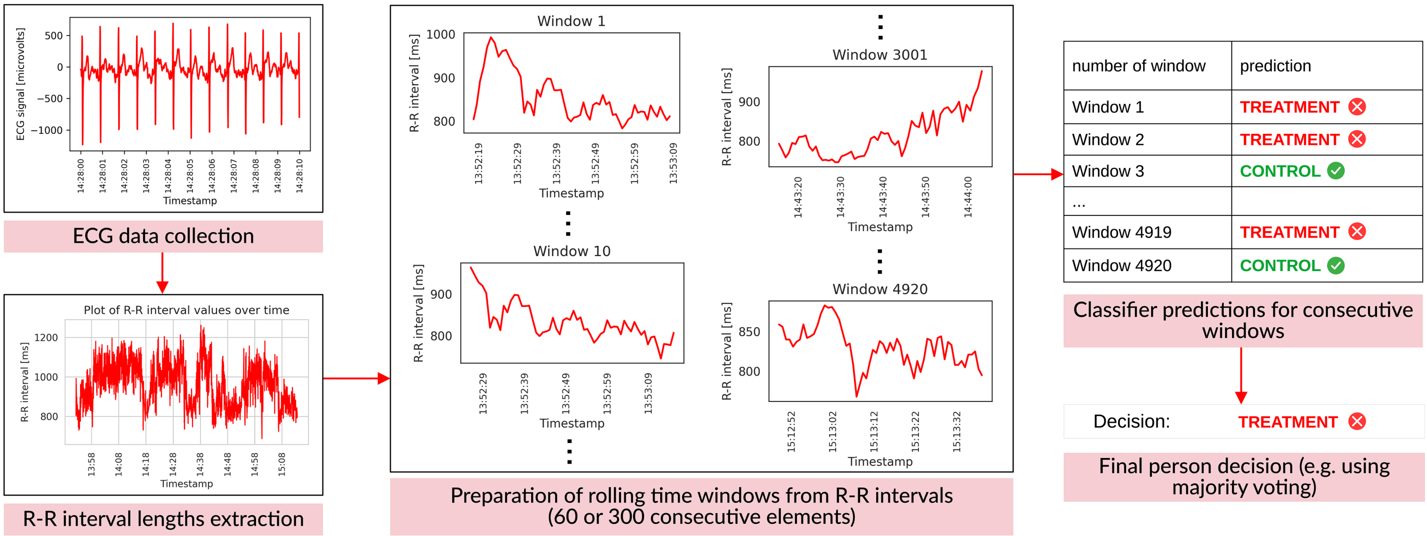
In this article, we study a classification approach for short-duration R-R interval windows to differentiate the data as belonging to a patient with schizophrenia/BD or a control group. R-R intervals are directly extracted from the ECG signal and various HRV metrics are based on their values [39,47]. Therefore, R-R intervals are a rich source of knowledge about ANS functioning and their consecutive values can be used as input features for classifiers.

Our main contribution is the assessment of the accuracy of a wearable-recorded classification-based diagnosis protocol. In reference to previous works, our specific contributions are as follows:

1. Larger data sample analyzed (60 participants vs 24/28 in previous works reporting classification results), distributed in a publicly available dataset;
2. State-of-the-art classification test scheme, including two levels of cross-validation, allowing for unbiased patient evaluation with automatic methods' hyperparameter optimization;
3. Comparison of several state-of-the-art machine learning classification approaches, including deep neural networks, allowing for their comprehensive evaluation, without hand-picked features.

Our approach can therefore be viewed as method-independent, providing a reference assessment that directly corresponds to a real-world clinical setting. This is particularly important as recent studies emphasize the need for advanced, automated ECG analysis techniques in various clinical applications [48]. The accuracy obtained (83% in the 5-fold cross-validation scenario and 80% in the leave-one-out cross-validation scenario) provides a solid, promising result for further applications. We view the application of our work as a prospect for the development of diagnostic and disease monitoring tools that combine traditional disease criteria based on observable symptoms [9,10] with physiological parameters such as the ECG signal for more accurate diagnosis [49]. Furthermore, we hope that in the future, with the implementation of biological marker-based machine learning-supported diagnostic methods, it will be possible to enhance predictive validity to assess the possibility of schizophrenia or BD in their prodromal state [50].

The outline of our method is presented in Fig 1 and will be discussed in detail in the next section.



number of window	prediction
Window 1	TREATMENT ⊗
Window 2	TREATMENT ⊗
Window 3	CONTROL ✓
...	
Window 4919	TREATMENT ⊗
Window 4920	CONTROL ✓

Fig 1. The scheme of the presented approach. First, for each person, ECG data is collected, from which a sequence of R-R interval lengths is extracted. Next, we apply a rolling window of size 60 or 300 to the aforementioned sequence of R-R interval lengths. The step between two consecutive windows is one, therefore there is an overlap of 59 or 299 R-R values between the two consecutive time series extracted by our rolling window algorithm. Then, for each window, a classifier predicts whether it belongs to the person from the control or treatment group. Finally, a single decision for a given person is returned, based on predictions for multiple windows.

https://doi.org/10.1371/journal.pcbi.1012983.g001

2. Methods

2.1. Ethics statement

All participants were informed and gave their written consent prior to admission to the experiment. The experimental procedure in the study received approval from the Bioethics Committee of the Medical University of Silesia, No.: BNW/NWN/0052/KB1/135/I/22/23.

2.2. Dataset

The dataset used for the classification experiments contains 1-1.5 hour recordings of R-R intervals for 60 participants, collected with a Polar H10 ECG chest strap suitable for measurements under different activity conditions [36]. Recordings were conducted during daytime hours, primarily around midday. The histogram illustrating participant activity is presented in Fig 2A. Out of the 60 participants, 30 belong to the control and 30 to the treatment group. The age range is 20 to 69 years in the treatment group and 24 to 69 years in the control group. In the patient group, 23 have been diagnosed with schizophrenia and 7 with bipolar disorder. Qualified psychiatrists diagnosed all study participants according to the International Classification of Diseases (ICD-11) guidelines. We decided to analyze schizophrenia and BD together as both are psychotic disorders with similarities in underlying mechanisms and HRV changes. For further explanation, please refer to Sect 4.1. The sex structure in both groups is comparable, i.e. the treatment group consists of 16 women and 14 men, while the control group includes 17 women and 13 men. The age distributions across the two compared groups are depicted in Fig 2B. All study participants, both in the treatment and the control group, are unique. The dataset is publicly available in [51].

Based on the analysis depicted in [39], all participants were examined using the Positive and Negative Syndrome Scale (PANSS) questionnaire designed for the assessment of symptom severity in psychotic disorders [52]. The mean values of consecutive PANSS subscores equal 16.37, 19.80, 35.30 and 71.47 for positive, negative, general and total scores, respectively. Furthermore, as shown in Fig 13 in [39], HRV was negatively correlated with the general score of PANSS ($r = 0.45$, $p = 0.01$). A more detailed evaluation of this dataset was performed in [39].

To perform an unbiased classifier evaluation, we divided the dataset into five separate folds in a standard cross-validation scheme, with windows containing a maximum of 60 or 300 consecutive R-R interval lengths, which correspond approximately to 1- or 5-minute recordings. Therefore, we have a total of 545604 or 521204 individual time-rolling windows with step length 1, respectively. Thus, there is an overlap of 59 or 299 R-R interval lengths between two consecutive time series.

During each of the five cross-validation steps, one fold is selected as a test set, while the remaining four folds are used for training and validation sets. A single fold contains all windows of six patients and six control individuals. This approach prevents data leakage between folds, as all measurements of a given individual are confined to a single fold. Training is performed in two phases: In the first phase, hyperparameters are optimized, while in the subsequent phase, a model is trained with the appropriate hyperparameter values. In the initial phase, during hyperparameter optimization, one of the four training folds is designated as the validation set. Once the optimal set of hyperparameters is determined, the four training folds are combined. During the final training phase, the data split into training and validation sets (within the four training folds) varies depending on the method, while the test set remains consistent across all methods.

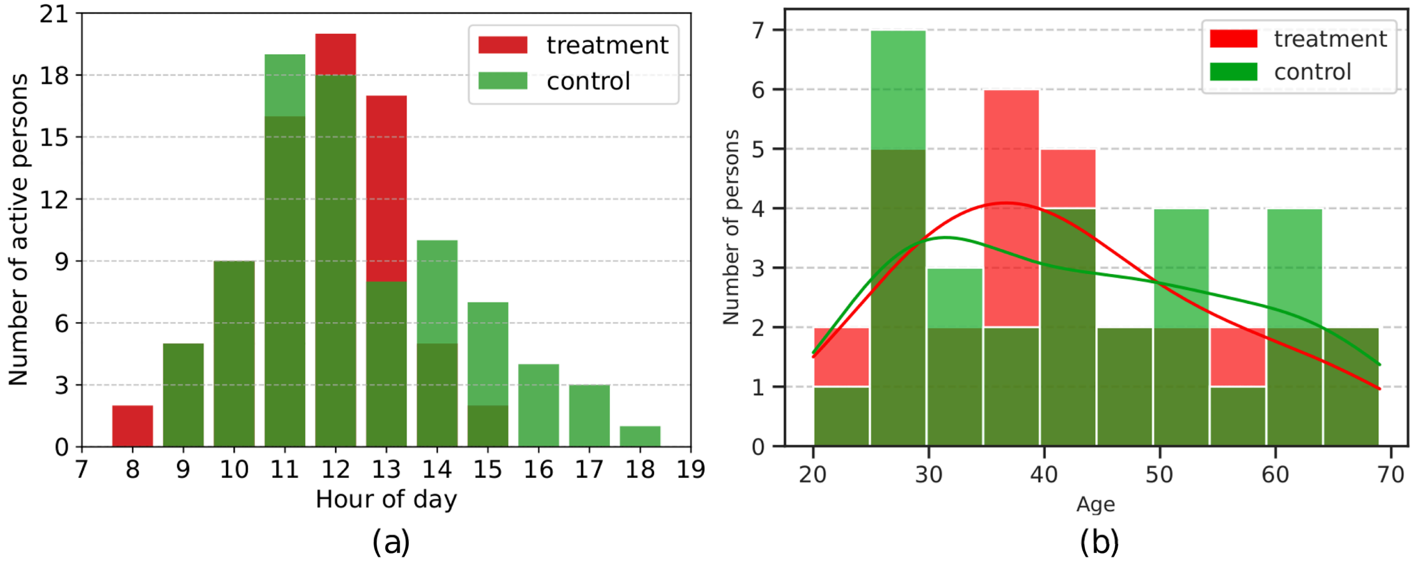


Fig 2. (a) Histogram of hourly activity for each group. Each bar indicates the number of individuals whose measurements covered the corresponding hour. (b) Histogram of age distributions across the control and treatment groups with kernel density estimation of ages.

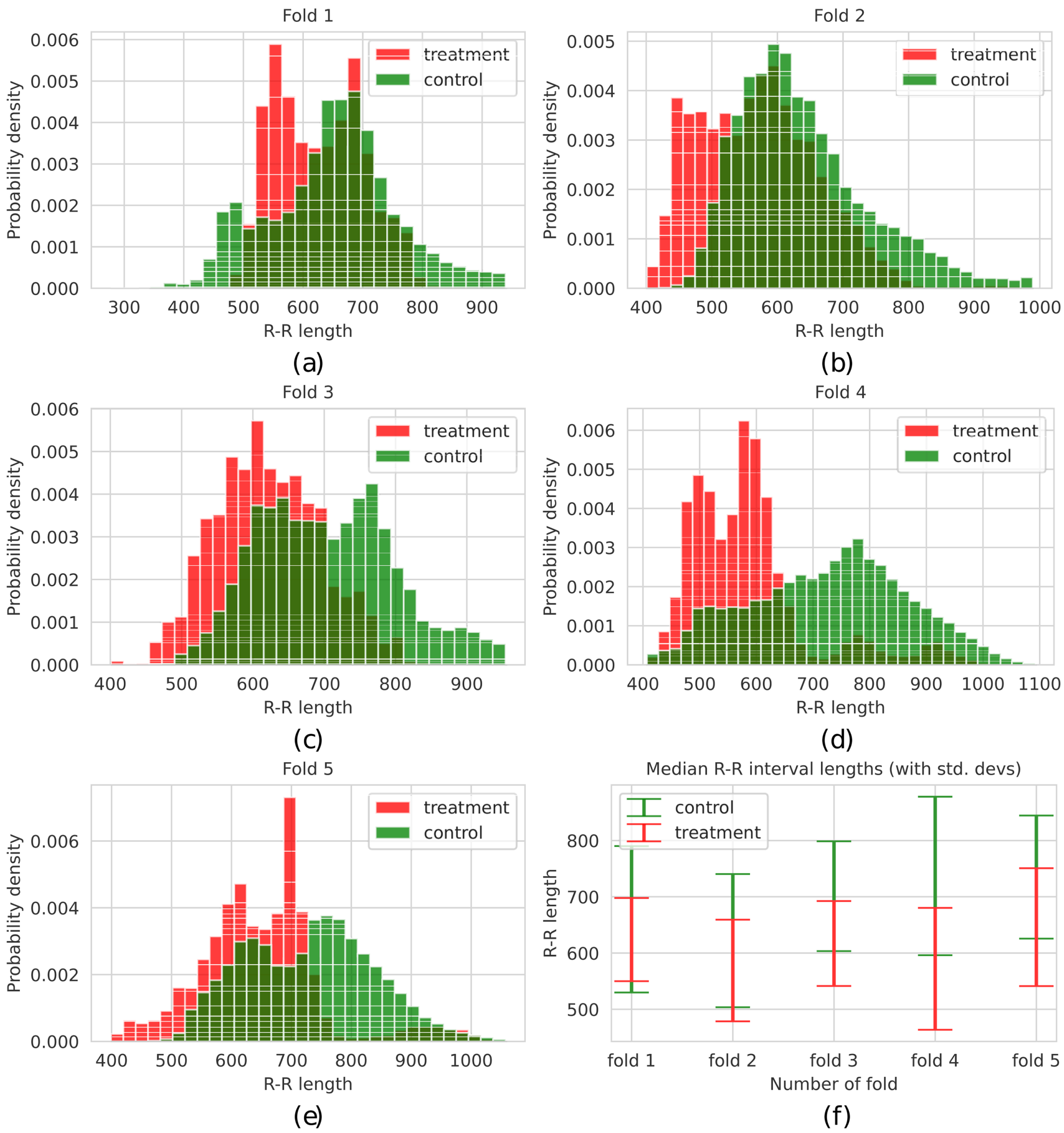
https://doi.org/10.1371/journal.pcbi.1012983.g002

For clarification, in Fig 3A, 3B, 3C, 3D and 3E, we depicted the distribution of individual R-R intervals' lengths for the treatment and control groups for five consecutive folds with 60-element time windows (consisting of approximately 1-minute long measurements), as well as median R-R values with the corresponding standard deviations in Fig 3F.

The presented plots reveal variations in data distributions across the folds. For example, fold 4 includes lower individual R-R values from the treatment group and higher values from the control group compared to the other folds. Fold 1 is different from the other folds in the sense that the distribution of the treatment group has two modes, and it largely overlaps with the distribution of the control group. This heterogeneity poses challenges for machine learning models; however, such distributional differences are often unavoidable in real-world scenarios involving limited data. To prevent selection bias and maintain the integrity and fairness of the experiments, we deliberately avoided manual adjustments in participant assignment across folds. Instead, individuals were randomly assigned to folds.

Our approach handles a real-world scenario in which long-term monitoring of many patients is challenging because patient signals have various characteristics. Performing experiments without severe limitations for evaluated persons (e.g., being in an isolated environment) is easier to arrange in practice but harder to analyze further. However, 60 persons with 1-1.5 hour recordings deliver a large number of individual time windows. We anticipate that brief events that alter the lengths of the R-R intervals, such as short walks, as described in the experiment procedure in [39], will not prevent the classifier from making accurate diagnoses for a substantial number of windows.

We also performed an additional leave-one-out cross-validation experiment to verify the impact of a larger and more diverse training set on the test results. In this scenario, recordings from 59 people form the training set during each training run, and the remaining person is in the test set. The procedure is repeated 60 times, each time for a different test person. This experiment is considerably more time-consuming than the 5-fold cross-validation and is therefore used solely as a supplementary analysis.



(a) (b) (c) (d) (e) (f)

Fig 3. The distributions of five folds designed for the cross-validation classification experiments with 60-element time rolling windows (a–e). Each fold contains six patients and six persons from the control group, and when merged, they form the whole dataset. In consecutive iterations, samples (sequences of R-R interval lengths) from four folds form a training and a validation set, while the remaining fold constitutes a test set. (f) Comparison of median R-R interval values with corresponding standard deviations across the two experiment groups and five data folds.

https://doi.org/10.1371/journal.pcbi.1012983.g003

2.3. Classifiers

From the theoretical point of view, our task, the diagnosis of schizophrenia or bipolar disorder based on the R-R interval lengths, can be seen as a time series classification task, for which various approaches have been introduced, including methods based on neural networks [53], Bayesian networks [54], hidden Markov models [55], genetic algorithms [56], Support Vector Machines (SVMs) [57], decision trees [58], frequent patterns (known as *motifs* or *shapelets*) and hubness-aware classifiers [59]. In our study, we considered the most prominent approaches that will be described in the following sections.

2.3.1. Nearest neighbor.

In [60], authors compared various time series classifiers, and concluded that 1-nearest neighbor “is an exceptionally competitive classifier”. For this reason, it has been widely considered a baseline in the last decades. Although the nearest neighbor classifier had been used mainly with dynamic time warping (DTW) distance, it was shown that the Euclidean distance may lead to similarly accurate results [61]. In particular, with increasing dataset size, the difference between the accuracy of the nearest neighbor with DTW and Euclidean distance is decreasing, and in the case of moderately large datasets, the difference is negligible [62]. At the same time, the Euclidean distance can be calculated in orders of magnitude faster than DTW, especially if the calculations are performed on GPUs.

For these reasons, we considered 1-nearest neighbor with Euclidean distance as a baseline in our experiments. For computational efficiency, we randomly selected 100000 training time series and searched for the nearest neighbors of test/validation time series only among those selected time series. Additionally, in line with the experiments with other models, this allowed us to repeat the experiments with various seeds that led to various selections of the training time series. The only hyperparameter we learned using the validation data was the decision threshold for classifying persons, i.e., the ratio of positively classified segments, so that the person will be classified as positive.

2.3.2. Multilayer percepton.

Multilayer perceptrons (MLP), i.e. fully connected feed-forward neural networks, with at least one hidden layer and non-linear activation, are known to be universal function approximators [63]. Therefore, MLP served as a deep learning baseline in our experiments. In particular, we considered MLP with a single hidden layer containing 128 units. We used the sigmoid activation function in the hidden layer and softmax activation in the output layer. We trained the model for 100 epochs using the Adam optimizer with cross-entropy loss. We learned the decision threshold for window classification, the decision threshold for person classification, as well as the most important hyperparameters of the optimizer, i.e., the learning rate and batch size, using the validation data. In particular, in each round of cross-validation, we performed a grid search and we searched for the best thresholds from $\{0.01, 0.02, \ldots, 0.99\}$, learning rate from $\{10^{-3}, 10^{-4}, 10^{-5}\}$, and batch size from $\{256, 512, 1024\}$.

2.3.3. SVM + feature selection with automatically configured hypothesis tests (FSH).

The method follows a classical approach to feature engineering with feature extraction and selection based on hypothesis tests. Initially, feature extraction is conducted using the “efficient” set of *tsfresh* [64] features. The set contains over 700 features, including statistical moments (mean, variance, skewness, and kurtosis), autocorrelation and partial autocorrelation, frequency domain features (Fourier/wavelet coefficients, spectral entropy), as well as linear model coefficients, distributional and quantile-based statistics and metrics related to peaks and changes. The overview can be found at [65].

Subsequent feature selection involves eliminating features with low variance and statistically identifying the 50 most pertinent features via the Mann-Whitney U test. Preprocessing is performed through robust normalization and random subsampling, in which the classifier is trained on a balanced subset of 1500 examples from each class. The classification task is executed utilizing an SVM with a Radial Basis Function (RBF) kernel. Parameter optimization for the SVM’s C and gamma hyperparameters is achieved through a grid search in the range from 10^{-1} to 10^{2}, based on the model’s performance on the validation set.

The threshold estimation for person classification is approached heuristically, determining it as the midpoint of the median scores across the classes derived from the training and validation sets.

2.3.4. Ensemble of SVMs. The ensemble classification scheme for R-R interval data follows the bagging paradigm. It utilizes a set of SVM classifiers $E = \{\mathcal{C}_i\}_{i=1}^{n}$, each trained on a different and exclusive subset of training data with hyperparameters C, γ determined by the method described in [66]. Those classifiers then perform a majority vote classification on each window in the test set. The label being a result of most of the SVM classifiers in E is treated as a result of E.

Then, a threshold t is calculated based on each person in the validation set data. For each person in the validation set, a R_i percentage of windows classified as 'i' is calculated, where $i \in \{0, 1\}$. Thus, R_i's are cleared of outliers, understood as data more distant from the mean than 2 standard deviations. Finally, the threshold t is set as the mean between the average values in R_0 and R_1.

The final classification of a person having the labels for each window assigned by E is obtained by determining whether a percentage of windows classified as class "positive" exceeds the threshold t.

Data preprocessing is performed for each window w by extracting features: $\min(w), \max(w), \mathrm{var}(w), \mathrm{mean}(w), \mathrm{median}(w)$.

2.3.5. XGBoost. XGBoost [67] or Extreme Gradient Boosting, is a scalable and efficient implementation of gradient boosting that has proven highly effective in various machine learning competitions and tasks. XGBoost, like the other tree-based methods, is very prone to overfitting, thus initially we used grid search to determine the hyperparameters. Two hyperparameters were checked, *NEstimators* and *MaxDepth*. The *NEstimators* parameter specifies the number of gradient-boosted trees used in the model, directly influencing the complexity of the model and its potential to overfit or underfit the data. The *MaxDepth* parameter controls the maximum depth of the trees, affecting the model's ability to capture complex patterns and its risk of overfitting with deeper trees. Using full cross-validation for each set of hyperparameters, we trained the model on $NEstimators \in \{10, 50, 100, 200, 500\}$ and $MaxDepth \in \{5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$. Then, we used the accuracy of the validation set to select the optimal hyperparameters. The optimal hyperparameters were $NEstimators = 200$ and $MaxDepth = 5$ in both cases, i.e. for 60 and 300-element time series. In the second stage, we used the validation subsets to determine the percentage of "positive" predictions that correspond to the highest person accuracy. If more than one value corresponded to the highest accuracy, we chose the value closest to 50%. The determined values were 49% and 44%, for 60 and 300-element windows, respectively.

2.3.6. GRU + FCN. Another deep learning architecture selected for the classification experiments was GRU + FCN, i.e. Gated Recurrent Unit and Fully Convolutional Network hybrid model [68], designed for univariate time series classification. The model architecture consists of two independent paths, the first of which is created by GRU blocks followed by the dropout layer, while the second part is formed by three blocks of convolutional layers followed by batch normalization and the ReLU activation in each of the Convolutional Neural Network (CNN) blocks. At the end of the second path, a global average pooling is applied. Finally, the output of the two model parts is merged and a fully connected layer assigns a label to a given sample. The authors of GRU + FCN proposed its approach based on Long Short-Term Memory+FCN (LSTM+FCN) [69], replacing LSTMs with GRU cells. We decided to include a recurrent neural network architecture in our studies because such an approach is well-suited for time series analysis, capturing short- and long-range dependencies.

For both 60- and 300-second time windows, we trained models through a maximum of 30 epochs with early stopping (based on the validation loss value), 10 epochs of patience and a learning rate scheduler with 5 epochs of patience. After that, the learning rate was reduced by an order of magnitude. We performed optimization of the crucial hyperparameters using the cross-validation procedure described in Sect 2.2. We searched for an initial learning rate from the set $\{10^{-4}, 10^{-3}\}$, batch size among $\{64, 128\}$, the number of GRU layers $\{1, 2\}$, the hidden state sizes $\{100, 200\}$, dropout probability (in the GRU path) $\{0.4, 0.8\}$, as well as for uni- and bidirectional recurrent networks. Furthermore, for the 60-element time windows, a batch size of 16 was evaluated.

As the highest-performing setup, we selected a network having a single layer of the unidirectional gated recurrent unit followed by a dropout layer with a probability of zeroing out values set to 0.8. Also, the convolutional layers were adjusted to 128, 256 and 128 kernels. The kernel sizes were set to 7, 5 and 3, respectively. For 60-element time windows, we selected an initial learning rate of 10^{-4}, a batch size of 128 and a hidden state size in GRU equal to 100. However, for the larger, 300-element time windows, an initial learning rate was about an order of magnitude higher than previously, that is, 10^{-3}, a batch size was reduced to 64 and a hidden state of GRU was extended to 200.

3. Results

3.1. Cross-validation experiments

In Table 1, we present results for all the evaluated machine learning methods in the case of 5-fold cross-validation experiments. As evaluation metrics, we selected window and person accuracies. More specifically, we compared the accuracy of all consecutive time windows with the accuracy for a given person based on a criterion specific to the method, e.g. majority voting. It is necessary to emphasize that not all windows for an evaluated person must be correctly classified to determine whether such an individual belongs to the control or treatment group.

We compared two variants of the time window length, i.e. 60 and 300 consecutive R-R interval lengths in a single window. Analysis reveals that the two considered time window length variants yield comparable performance, suggesting that providing additional contextual information, such as extending measurements to approximately five minutes, is unnecessary for the classifiers. This conclusion is important from a practical point of view because

Table 1. Experimental results for different methods and time window lengths averaged over 5 test folds, in terms of overall accuracy [%]. Also, in the case of the non-deterministic methods, results are additionally averaged over 5 training runs per each cross-validation step. Bold font indicates methods with the highest accuracy in a given category. MLP – multilayer perception; FSH – feature selection with automatically configured hypothesis tests; GRU – Gated Recurrent Unit; FCN – fully convolutional network.

method	windows length	test window accuracy	test person accuracy
Nearest neighbor	60	65.62 ± 5.4	77.33 ± 4.4
	300	58.08 ± 6.4	70.00 ± 12.5
MLP	60	**74.10 ± 7.4**	80.67 ± 8.7
	300	70.99 ± 6.7	73.33 ± 9.7
SVM + FSH	60	73.37 ± 7.2	80.33 ± 10.2
	300	**73.65 ± 6.7**	79.00 ± 7.3
Ensemble of SVMs	60	73.74 ± 5.0	81.67 ± 6.2
	300	73.54 ± 4.9	**83.33 ± 5.3**
XGBoost	60	72.22 ± 8.3	**83.33 ± 8.3**
	300	71.45 ± 8.7	80.00 ± 12.6
GRU + FCN	60	69.97 ± 8.1	80.00 ± 11.7
	300	71.75 ± 11.0	79.67 ± 13.2

https://doi.org/10.1371/journal.pcbi.1012983.t001

300-element time windows are more computationally expensive while not bringing clear advantages. Also, most of the methods perform even slightly better for shorter windows.

We also noticed that the methods based on the ensemble of classifiers like XGBoost and Ensemble of SVMs lead to the highest test person accuracy, exceeding 80% for both compared window settings. Taking into account that the final person's decision may be based on multiple time windows, this experimental scenario is more favorable for ensemble methods. They are designed to benefit from the variety of information contained in different windows. Their performance is the highest, even though the accuracy for individual windows is slightly higher for MLP and SVM + FSH. However, it is necessary to emphasize that the differences between the compared methods are not huge and, in general, they achieve similar accuracy, except for the nearest neighbor classifier having the lowest scores in each category. Nevertheless, in such an experimental setting, which corresponds to the real-world scenario with limited possibilities of patient monitoring, we can still classify 8 out of 10 evaluated persons correctly, on average.

The results of the statistical tests confirm these observations. For both considered window lengths and two evaluated performance metrics, we applied the Friedman test to compare multiple classifiers. In the case of 60-element test windows, the Friedman test statistic for per-window accuracies was 11.97 with a corresponding p-value of approximately 0.035, indicating statistical significance at the 0.05 level. For per-person accuracies, the test statistic was 3.63 with $p \approx 0.6$, showing no significant differences.

Following the significant finding for per-window accuracies, we conducted pairwise comparisons using the Wilcoxon signed-rank test with the Benjamini-Hochberg correction for multiple testing. However, none of the pairwise differences reached statistical significance. To further examine the magnitude of differences between classifiers, we computed Cohen's *d* for paired samples, as presented in Fig 4A. These effect sizes confirmed a considerable discrepancy between the nearest neighbor classifier and the other methods, with *d* values ranging from 0.81 to 2.14.

We performed an analogous analysis for 300-element time windows. The Friedman test yielded a statistic of 11.63 ($p \approx 0.04$) for per-window accuracies, and 6.20 ($p \approx 0.29$) for per-person accuracies. Post-hoc Wilcoxon signed-rank tests, adjusted with the Benjamini-Hochberg procedure, again indicated no statistically significant differences between pairs of classifiers. Nevertheless, Cohen's *d* values, depicted in Fig 4B, pointed to consistently large discrepancies between nearest neighbor and the remaining classifiers, albeit with smaller effect sizes for other pairs of classifiers compared to the 60-element time windows. However, we note that these findings are biased with a relatively small sample size, since results were averaged over five folds per method.

We performed an additional series of experiments in a leave-one-out cross-validation scenario for two methods, i.e. Ensemble of SVMs and GRU + FCN. Their results are presented in Table 2 while the distributions of consecutive in-person accuracies for 60 test folds are depicted in Fig 5A and 5B. The two analyzed methods achieved comparable performance, both in terms of test window accuracy (70.1% for Ensemble of SVMs and 71.8% for GRU + FCN) as well as for test person accuracy (80% for Ensemble of SVMs and 79.4% for GRU + FCN). Based on the presented histograms, we can observe that for the majority of persons, more than 50% of windows were classified correctly. Also, the median fold accuracy was equal to 76.3% for GRU + FCN and 77.9% for Ensemble of SVMs.

Additionally, the statistical comparison of the two considered approaches revealed no significant differences in either per-window or per-person accuracies. The Wilcoxon signed-rank test statistics were 851 ($p \approx 0.64$) and 121.5 ($p \approx 0.86$), respectively.

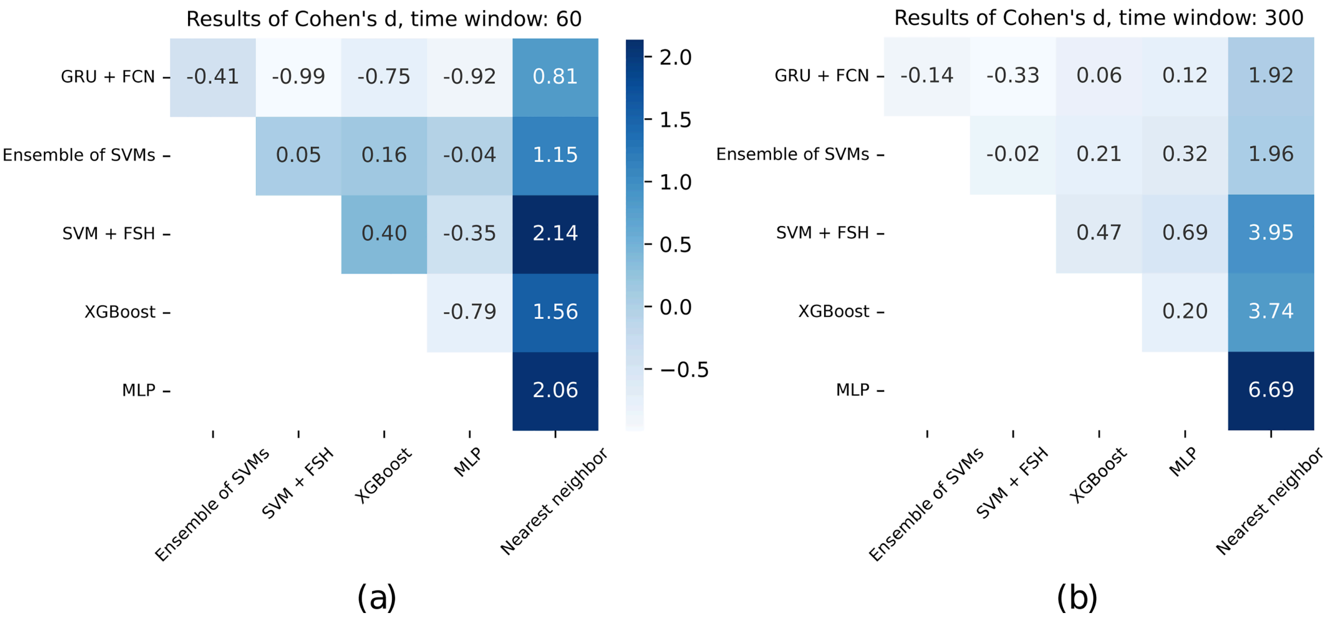


Fig 4. Results of Cohen's *d* between consecutive pairs of classifiers, calculated for the two considered time window sizes: 60 (a) and 300 (b), corresponding to the experiments described in Table 1.

https://doi.org/10.1371/journal.pcbi.1012983.g004

Table 2. The results of the leave-one-out cross-validation experiment averaged over 60 test folds (in terms of overall accuracy [%]). In the case of the non-deterministic method, like GRU + FCN, results are additionally averaged over 3 training runs for each cross-validation step. Bold font indicates the highest-performing method in each category.

method	windows length	test window accuracy	test person accuracy
Ensemble of SVMs	60	70.08 ± 26.5	**80.00 ± 40.0**
GRU + FCN	60	**71.75 ± 22.2**	79.44 ± 37.6

https://doi.org/10.1371/journal.pcbi.1012983.t002

3.2. The impact of data features on classifier decision

To analyze the influence of features on model decisions, we employed SHapley Additive exPlanations [70] (SHAP). SHAP quantifies each feature's contribution to a model's prediction (here, the probability of belonging to the positive class), ensuring that, for every instance, the sum of SHAP values for all features equals the difference between the model's output and its baseline (average) prediction.

For this investigation, we have manually selected a small set of seven distinct, well-known and intuitive features: mean, minimum, maximum values, standard deviation, skewness, kurtosis, lag-i autocorrelation function (ACF) values in each window, where $i \in \{1, 2, 3, 4\}$. We then applied a Random Forest (RF) classifier to a dataset resampled by selecting every 500th instance (so the visualization is not overloaded). In a standard 5-fold cross-validation experiment performed on this subsampled dataset, the RF classifier achieved 73.3% accuracy with 60-element windows and 73.9% with 300-element windows, suggesting that these features are effective data descriptors. The importance of the features obtained was similar in both experiments, so only the 300-element windows' experiments are presented here.

Fig 6 presents the bee-swarm plot that highlights the distribution of the impact of the features on model predictions and shows how this impact correlates with the feature values

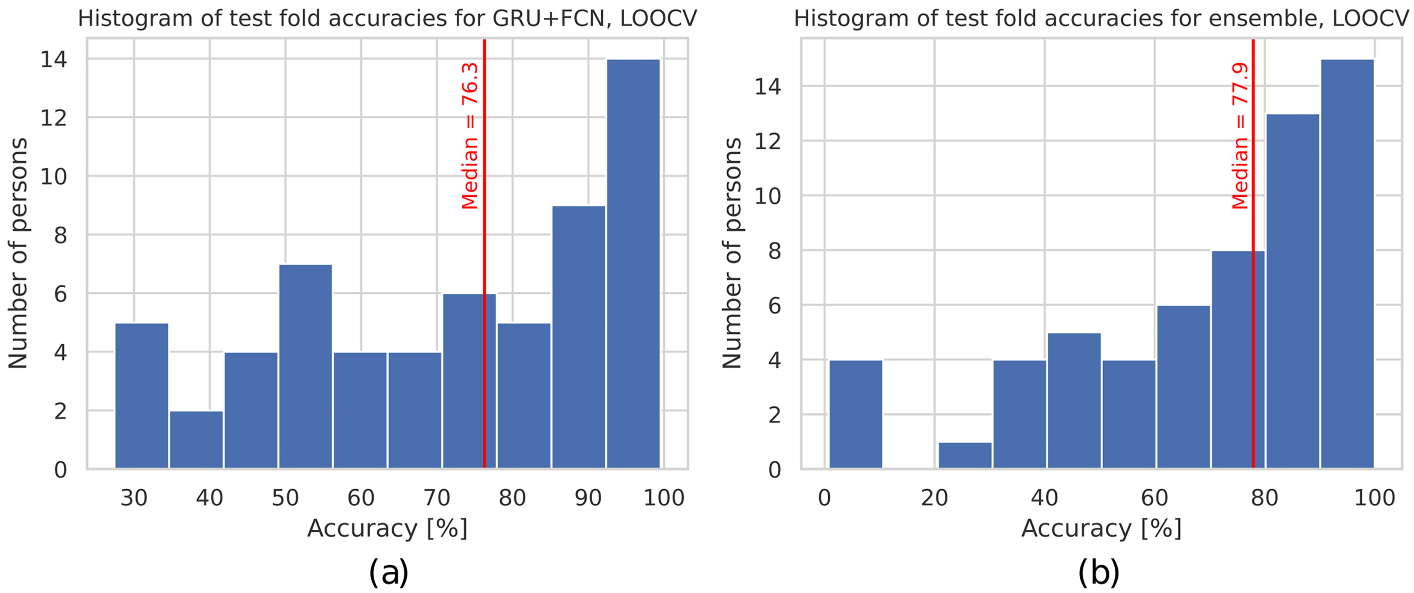


Fig 5. The distributions of 60 test fold accuracies for the leave-one-out cross-validation experiment for GRU + FCN (a), averaged over three different runs, and Ensemble of SVMs methods (b).

https://doi.org/10.1371/journal.pcbi.1012983.g005

(high/low). It reveals that standard deviation and maximum values are the most essential features, with high values (red) linked to the negative class (control group), and low values (blue) related to the positive class (treatment group), especially for standard deviation. This tendency agrees with our observations that the low variance of R-R values and shorter beat-to-beat intervals are characteristic of the treatment group.

Lag-i ACF values, where $i \in \{1, 2, 3, 4\}$, show relationships of the currently considered R-R interval to their first, second, third and fourth neighboring R-R interval values, respectively. These variables are considerably important in terms of influencing the model predictions. High lag-4 and lag-3 ACF values are associated with the positive (treatment) group, while high lag-1 and lag-2 ACF values are related to the control group.

We extended our analysis concerning ACF and separately calculated its values for all data instances, as presented in Table 3. The highest autocorrelation is observed between two consecutive R-R intervals. However, its mean values are higher for the control than for the treatment group, especially for the 60-element time series (0.84 ± 0.1 vs 0.78 ± 0.2). This tendency reverses for larger lag values, and finally, for lag-4, ACF is much higher for the treatment group (0.74 ± 0.17 vs 0.61 ± 0.21 for 300-element time series) than for controls. These observations are further discussed in Sect 4.3.

Kurtosis and skewness have a visibly lower impact on the classifier decision, which may result from the fact that R-R intervals, especially within a short window, are likely to have a relatively symmetrical distribution if the heart rate does not change significantly due to external factors.

Fig 7 presents the heatmap plot where the X-axis corresponds to the instances ordered by the model output value (sum of SHAP values) shown at the top of the plot. We can see how the examples for which the model is most confident in its decision are mainly impacted by a small subset of features, and how other features are more relevant for less obvious instances located in the middle of the presented X-axis range.

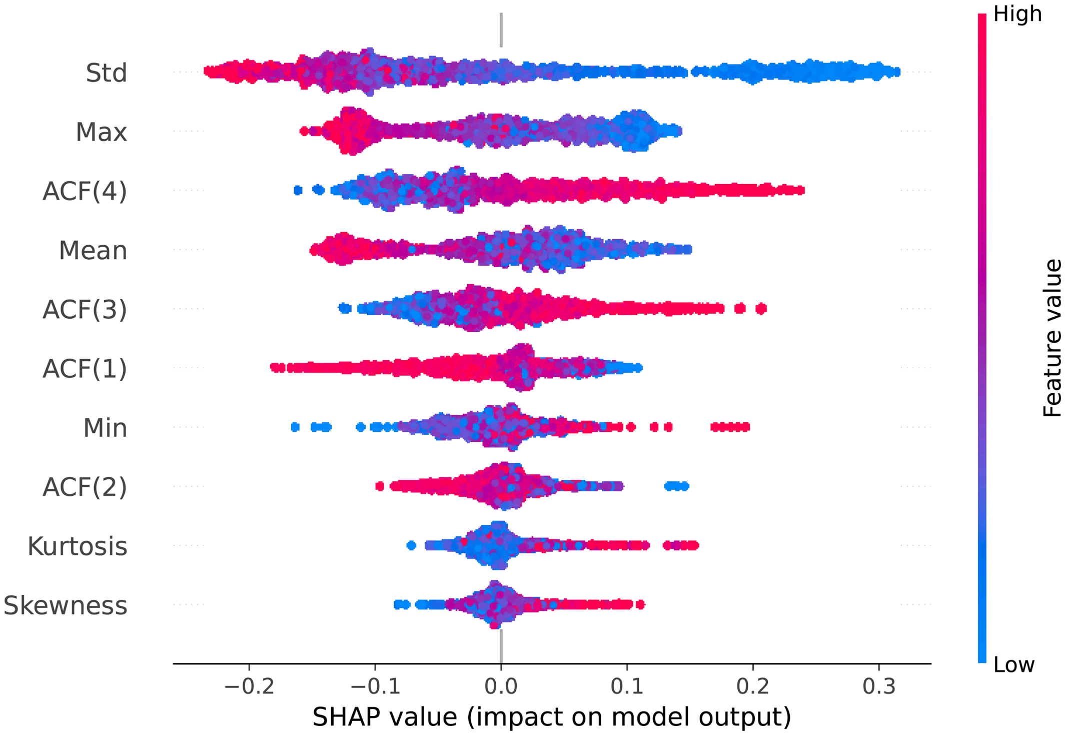


Fig 6. The distribution of feature impact on model predictions of the treatment (positive SHAP values) and the control group (negative SHAP values) as a SHAP bee-swarm plot. Every point represents a data instance with clusters indicating the data density. Rows and columns correspond to features and SHAP value, respectively, while color intensity represents the feature value. The red color indicates high values of the considered features, while the blue color represents the opposite case. Features are ranked by their impact on the model output. Data is from a 300-element windows' experiment on a dataset sampled every 500th instance.

https://doi.org/10.1371/journal.pcbi.1012983.g006

Table 3. Autocorrelation function with lag values, i.e. ACF(i), where $i \in \{1, 2, 3, 4\}$, for control and treatment groups across all time windows. For lags greater than 2, the mean ACF values for the treatment group consistently exceed those of the control group.

		mean ± standard deviation	
lag	window	control group	treatment group
ACF(1)	60	**0.84** ± 0.12	0.78 ± 0.20
ACF(1)	300	**0.91** ± 0.07	0.90 ± 0.11
ACF(2)	60	**0.67** ± 0.19	0.65 ± 0.25
ACF(2)	300	0.81 ± 0.12	**0.84** ± 0.13
ACF(3)	60	0.51 ± 0.23	**0.58** ± 0.24
ACF(3)	300	0.70 ± 0.17	**0.79** ± 0.15
ACF(4)	60	0.37 ± 0.27	**0.50** ± 0.24
ACF(4)	300	0.61 ± 0.21	**0.74** ± 0.17

https://doi.org/10.1371/journal.pcbi.1012983.t003

3.3. Classifier diversity and ensemble methods

Due to the diverse patterns among evaluated individuals, each scenario contains a subset of data that is misleading for classifiers. Although the evaluated methods are consistent in straightforward cases, as shown in Fig 8A, they generate numerous errors in ambiguous cases, resulting in a high number of misclassified time windows, like for the individual depicted in Fig 8B. The classifiers occasionally disagree, as presented for the individual in Fig 8C, for

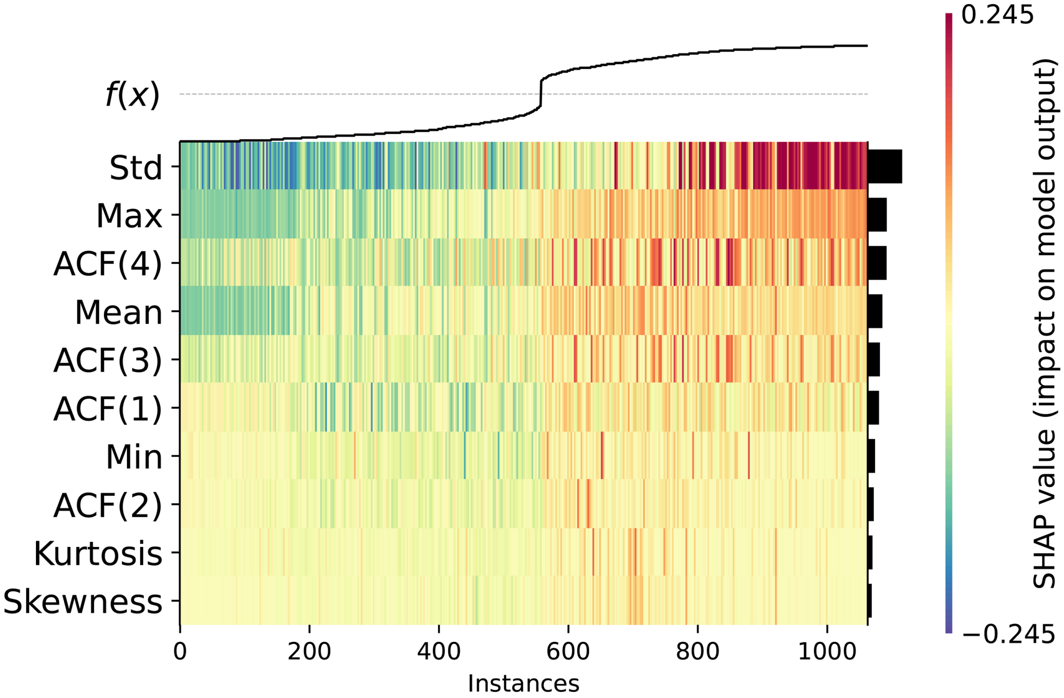


Fig 7. The influence of features on model predictions, presented as a SHAP heatmap. Rows and columns correspond to features and instances, respectively. Instances are sorted according to the model's output score. Color intensity represents the SHAP value, with red indicating a positive impact and green a negative impact on the model output (probability of belonging to the positive class) relative to the baseline. Features are ranked by their importance and are also presented as sidebars. Data is from a 300-element windows' experiment, on a dataset sampled every 500th instance.

https://doi.org/10.1371/journal.pcbi.1012983.g007

which GRU + FCN incorrectly classified most windows, whereas Ensemble of SVMs was correct in nearly all its predictions.

This observation aligns with the no-free-lunch theorem, which states that there is no universal classifier with high performance for diverse datasets. Although each evaluated method may focus on different patterns, even for correctly classified individuals, a subset of time windows is misclassified due to locally varying characteristics. This considers patients in both movement and rest conditions. For example, for control individuals, the R-R intervals decrease during activity and exhibit low variance within the analyzed windows, making them more similar to the individuals in the treatment group, who generally have shorter R-R interval lengths with slight variance (see Figs 7 and 9A).

Stacking classifier. The high level of disagreement between classifiers (see e.g. Fig 8C) raises the question of whether forming a stack of classifiers would improve classification accuracy. To investigate this, we conducted an additional experiment using a stack of three methods from architecturally diverse model families: GRU + FCN, Ensemble of SVMs, and XGBoost, with a logistic regression classifier. The classifier input to the logistic regression model was a three-element sequence of 0's or 1's based on the predictions of each method for a given time window. However, the final classification accuracy remained similar to the best result among the individual models, indicating that, in this case, the use of multiple methods does not significantly improve overall results.

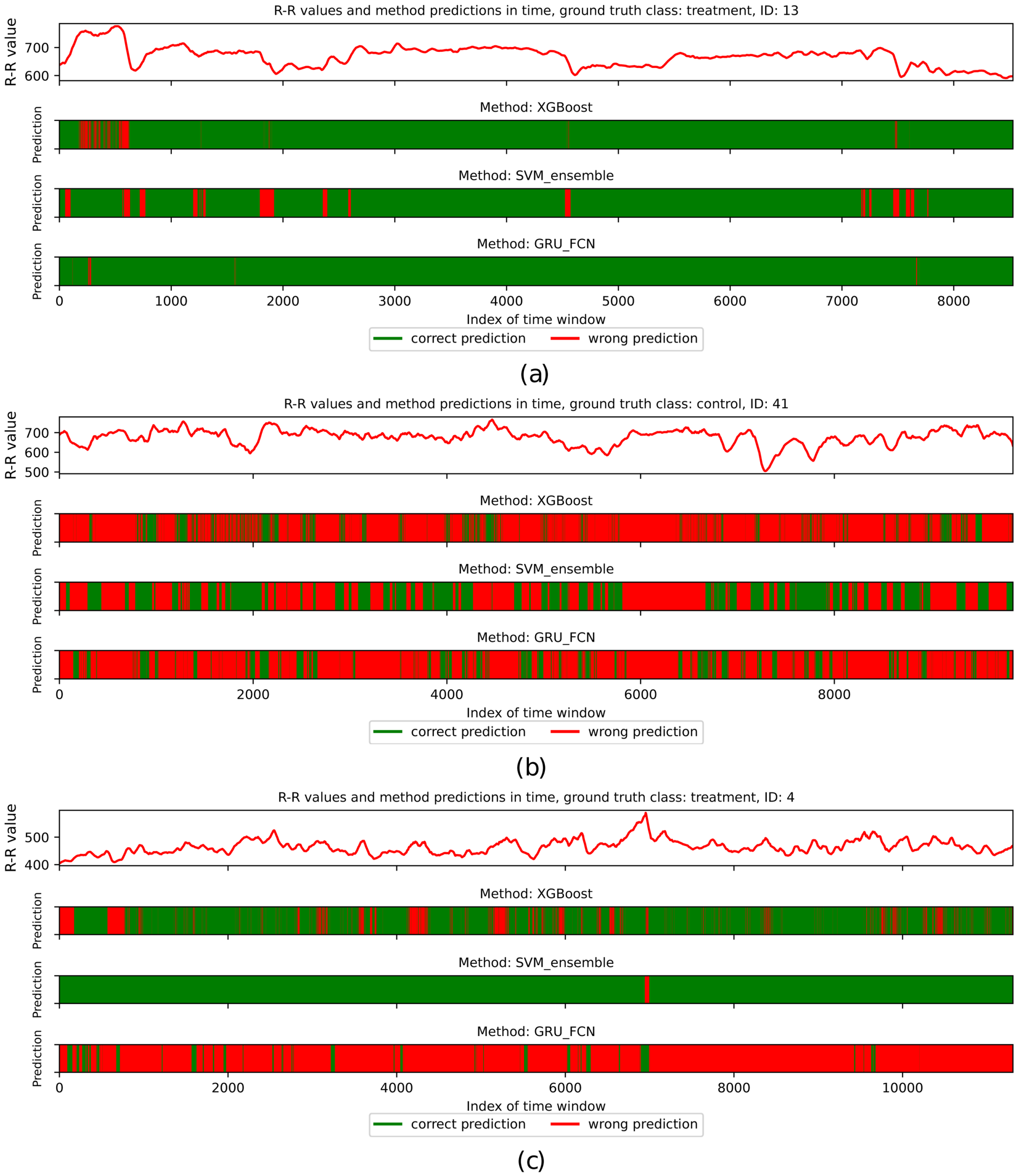


Fig 8. The overview of R-R values and predictions of XGBoost, Ensemble of SVMs, and GRU + FCN classifiers for consecutive time windows for the three selected individuals. Green areas correspond to the correctly classified periods, while red areas refer to the opposite case. (a) depicts the individual from the treatment group being relatively straightforward for all three classifiers. (b) corresponds to the selected control group individual for whom most of the time windows are classified incorrectly by all three compared methods. Finally, (c) represents the treatment group individual classified mostly accurately by XGBoost and Ensemble of SVMs but misclassified by GRU + FCN.

https://doi.org/10.1371/journal.pcbi.1012983.g008

Long-term predictions. The time windows in this study were approximately 1 or 5 minutes in duration, corresponding to 60 or 300 consecutive R-R intervals, respectively. Short-term measurements in randomly selected time windows may lead to incorrect predictions in changing conditions. However, in many cases, classification with 60-element time windows yielded higher accuracy than for the 300-element sequences, as presented in Table 1.

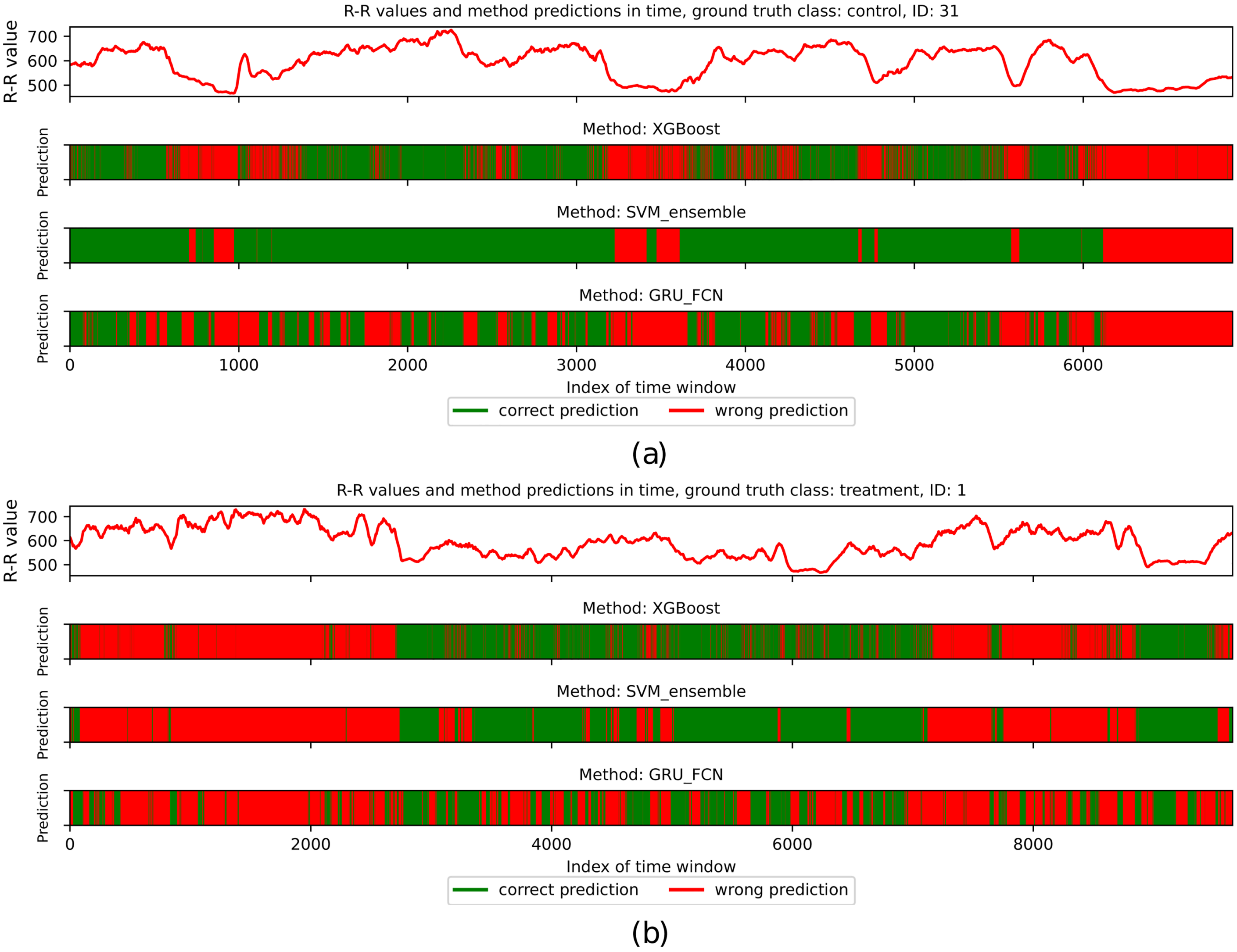


Fig 9. The overview of R-R values and predictions of XGBoost, Ensemble of SVMs, and GRU + FCN classifiers for consecutive time windows for the two individuals. Green areas correspond to the correctly classified periods, while red areas refer to the opposite case. (a) corresponds to the control group individual for whom selected contiguous time windows were classified incorrectly. All three tested methods made errors for periods corresponding to lower R-R interval lengths, while for the remaining time windows, the Ensemble of SVMs achieved the highest performance. (b) depicts the individual from the treatment group whose signal is demanding for classifiers. Only the middle part of the measurements mostly led to the correct labelling (except for GRU + FCN, making many prediction errors but still less than for the remaining parts of the signal).

https://doi.org/10.1371/journal.pcbi.1012983.g009

Therefore, extending the time range within individual windows seems to be unnecessary. Nevertheless, as shown in Fig 9A and 9B, some time periods may be easier to classify than others. Thus, the selection of windows merits consideration in future experiments and may increase the classification accuracy.

Minimum variance critertion. We evaluated an alternative voting strategy (compared to the one described in Sect 2.3) for assigning the final label which relies on identifying the most extended period of stable predictions characterized by the lowest prediction variance. This approach is based on the premise that consistent predictions over a long time period should provide more reliable information than predictions across varying states. To evaluate this, the variance was calculated based on 60 consecutive time window predictions. Then, the longest time range having the lowest variance was considered, and the most common class within this subset was set as a model prediction.

The experiment was carried out using the three methods described in Sect 2.3: Ensemble of SVMs, XGBoost, and GRU + FCN. This strategy improved the accuracy of XGBoost to 85.00 ± 12.25%, the highest result among all methods tested. However, since it did not improve the performance of the other methods, the approach was ultimately not adopted.

4. Discussion

4.1. Schizophrenia and bipolar disorder as a common group of psychotic disorders

Although schizophrenia and bipolar disorder are classified as two distinct diseases based on the phenomenological description of symptoms [9,10], there is an ongoing debate whether both diseases actually represent a continuum [71,72]. Based on MRI findings on the neurological level, both diseases share a common altered network of brain areas [71], both manifest as psychotic disorders [37] and both share many common genetic, endophenotypic and phenomenological traits [38]. One of such prominent endophenotypic traits is the shared difference in ANS regulation in comparison to healthy controls as observed through heart rate variability that is well described in both schizophrenia [21] and BD [73]. For these reasons, we chose to analyze both diseases collectively, aligning with other studies that examine these groups together with respect to severity of symptoms [32] or by broadly defining the studied population as psychotic [74].

4.2. The effect of antipsychotic medication on classification results

The influence of antipsychotic medications on ANS dysfunction is well established, with a diminishing ANS activity and subsequent change in HRV, especially in the patient group taking the antipsychotic quetiapine [75]. This raises the question whether the classifier results actually represent a distinction between psychotic patients and healthy controls, or rather a distinction between patients under antipsychotic medication with an impact on HRV and healthy controls.

It was not possible to directly resolve this question in our research by withdrawing patients' medication during hospitalization due to bioethical concerns, as this could have a major detrimental effect on the course of the disease. Although it is important to remember that it is already well established that psychotic patients generally have changes in ANS activity and HRV [21,73]. Recent work considering this dataset [39] indicated that HRV is related to the severity of psychotic symptoms, representing a global change in the activity of CNS (central nervous system) and ANS. In addition, patients in our group were treated with a range of antipsychotic medications. Given that different antipsychotics exhibit distinct pharmacological profiles and varying effects on the autonomic nervous system, the treatment group is likely to exhibit greater heterogeneity, potentially complicating classification efforts. It is also important to take into the account that in our research, patients and healthy controls performed various degrees of activity during a time-span of more than one hour. Considering that the influence of ANS activity on cardiac regulation is multifactorial and can be interpreted on many levels, such as in the neurovisceral integration model [76], the effect of antipsychotics should be even out with other factors that appear during normal activity, resulting in HRV regulation representing global ANS activity and not antipsychotic medications selectively.

However, to assess the potential effect of antipsychotic use on R-R intervals in our study, we compared the distributions of the R-R values between seven patients receiving quetiapine, which has the most significant diminishing effect on sympathetic and parasympathetic ANS activity [75] and the remaining participants in the treatment group. The histograms for both subgroups are presented in Fig 10. We expected that if the classifier results were influenced mainly by antipsychotic medication, the antipsychotic with the largest effect on ANS would result in a significantly different R-R distribution; however, no substantial differences were observed. This result is consistent with observations in [39] where no significant effect of quetiapine on mean HRV scores was identified compared to other antipsychotics. In general, our analysis does not reveal a significant impact of quetiapine compared to other antypsychotics on the distribution of R-R intervals.

Even if antipsychotic medications were proven to have an influence on the features utilized by the classifiers, this phenomenon does not undermine the importance of ongoing research efforts in this area. In such cases, these classifiers would, for example, help clinicians distinguish between patients who comply with treatment guidelines and those who have stopped taking their medicine and are at risk of disease remission.

4.3. Relationship between ACF values and ANS balance

Based on the SHAP values presented in Fig 6, we can observe that the high lag-3 and lag-4 R-R interval autocorrelation values are more common in the treatment group, while the high lag-1 and lag-2 autocorrelation are mostly related to the control group. This may be partially explained by the fact that parasympathetic and sympathetic regulation of HRV through the ANS occurs at different time intervals, with a maximum response of about 0.5 seconds for the parasympathetic branch and about 4 seconds for the sympathetic branch [77,78]. Given

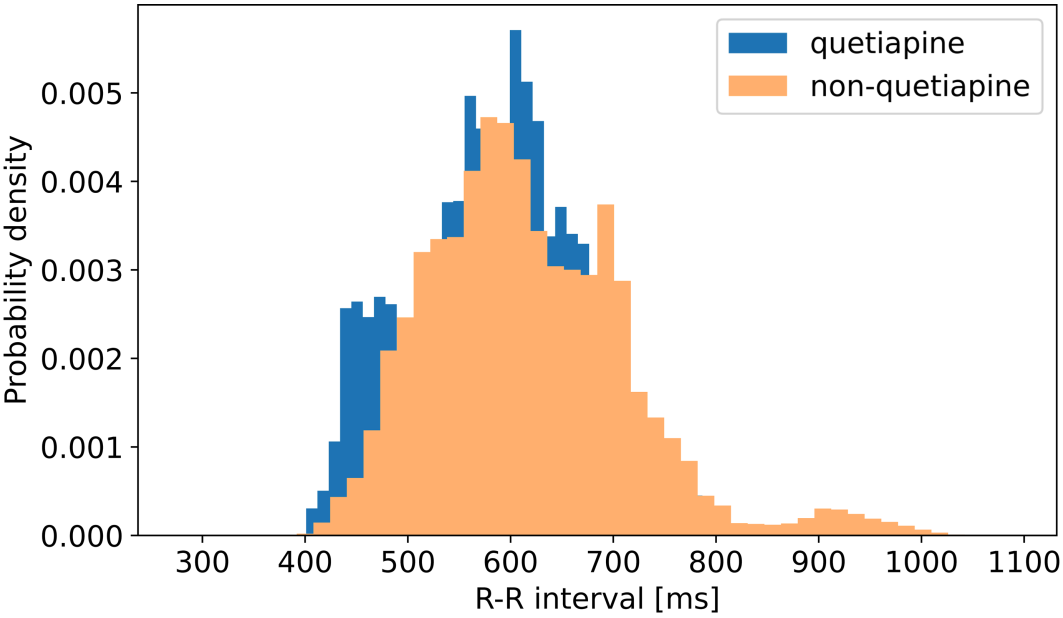


Fig 10. Histogram of R-R intervals for treatment group participants, stratified by quetiapine usage (quetiapine vs. non-quetiapine treatment subgroups).

https://doi.org/10.1371/journal.pcbi.1012983.g010 that the ACFs of successive lags correspond to the aforementioned time intervals, these features have the potential to distinguish between parasympathetic and sympathetic activity. This is in line with the fact that the beat-to-beat regulation of HRV represented by lag-1 is parasympathetic in its nature [79] and that psychotic patients have an impaired parasympathetic response [21,80,81]. It suggests that lag-1 remains an important feature for the control group due to the proper functioning of the parasympathetic system and beat-to-beat regulation. Consequently, parasympathetic dysregulation would shift the balance in favor of a stronger sympathetic response in the patient group [21,82], which could correspond to what we can observe as a growing strong relationship between the high value of lag-3 and the subsequent lag-4 feature and the prediction in the patient group. In this case, the lower value of lag-2 could represent the moment of transition moment between the peak parasympathetic and sympathetic response, in which neither the sympathetic nor parasympathetic branch of the ANS has a dominant impact on this characteristic. This hypothesis is strongly supported by changes in the distribution of the impact of the feature on model predictions between subsequent lag ACF values.

There is an ongoing debate on whether the ANS imbalance in favor of higher sympathetic activity in psychotic patients occurs mainly because they have higher baseline sympathetic activity [82] or mainly because they have lower parasympathetic activity [21]. As we mentioned earlier, the parasympathetic system has a short response latency of 0.5 seconds and a return to baseline within approximately 1 second [78]. Therefore, the influence of parasympathetic activity on lag-1 and possibly also lag-2 should be high, probably with little to no impact on lag-3 and even less on lag-4. In the case of our results, given that the impact of lag-3 and lag-4 ACF on the prediction of the model is greater than that of lag-1 and lag-2 ACF, this could suggest that psychotic patients not only have a sympathetic shift in ANS balance but mainly have a higher baseline level of sympathetic activity compared to the control group. In a scenario in which psychotic patients predominantly exhibit reduced parasympathetic activity, the influence of sympathetic activity, as reflected in lag-3 and lag-4 ACF values, would be expected to have a diminished impact on prediction within the patient group. Conversely, in the control group, the predictive contribution of lag-1 ACF, which represents parasympathetic activity, would be expected to be comparatively higher. However, this is not the case in our sample.

4.4. Relation of our study to other works

In this work, we have evaluated a diverse set of classifiers to generalize on particular characteristics of different algorithms. Additionally, our cross-validation scheme removes the bias of the manual choice of hyperparameters. Our experiments address a broad spectrum of key methods in ECG processing, as outlined in [48]. Together, we view this as a solid foundation for the quantification of the possible accuracy of a wearable HRV diagnostic system. The different methodologies employed by the authors of [41] and [42] complicate direct comparisons. However, our findings are broadly consistent with theirs and significantly enhance the generalizability of the evaluation of this diagnostic framework. Our paper supplements such works as [33] or [43], where no actual classification takes place.

When measuring HRV, an important variable is the length of the recorded sequence and patient activity. In this work, we use the dataset described in [39] which comprises relatively short time measurements (on the order of 1h) and lightly structured activity (patient-defined activity punctuated by walking). This limitation of time and broadly defined instructions are advantageous and provide flexibility (e.g. applicable to both on-site and at-home diagnostics) and provide minimal burden for the patient. We view this as a more accessible and versatile approach than long-term recording on the order of days [33,41] or structuring the activity [42].

The diversity of patients is an important factor to consider when evaluating a diagnostic scheme. The dataset we considered is significantly larger (60 persons) than in previous works reporting classification results (24 in [41] or 28 in [42]), while simultaneously maintaining similar or higher accuracy.

Our choice of sampling window (300 samples) corresponds to the theoretical analysis of short-term cardiorespiratory dynamics presented in [83], where it was found to provide a feasible assessment of cardiac dynamics variability from electrocardiographic records. Our experimental scheme is related to the one presented in [84], where three subsequent stress interventions were used, preceded by a baseline record and each followed by a rest period. In our case, walking interventions were used (see [39] for details) to achieve HRV tracking in a variety of activities.

4.5. Comparison of EEG- and ECG-based methods in psychiatric diagnostics

Most available research in the field of EEG in schizophrenia and BD focuses on the analysis of the electrical activity of the brain cortex in the resting state [85]. Consequently, machine and deep learning classification research in this field focuses mainly on resting state data [86,87], partially because some of these EEG datasets are publicly available [88]. Another branch of EEG research in schizophrenia and BD is the analysis of event-related potentials (ERP), which are short EEG segments that are linked to particular events of experimental interest and are repeated through numerous trials to filter out signal noise [89]. One such ERP of particular interest is mismatch negativity (MMN), which has a well-documented link to psychotic disorders [90,91]. Another ERP of potential use in psychiatry is the P300 component of auditory evoked potential, with some research showing its relationship to mood disorders, including BD [92]. ERP has potential as a tool for disease treatment and progression monitoring, mainly in the area of patient cognition evaluation [93].

Deep Learning (DL)- assisted whole-brain resting-state EEG activity has a high diagnostic potential, although limitations arise from its complexity and reluctance to adapt to clinical use by caregivers [86]. Machine and deep learning methods using EEG tend to have a higher accuracy rate than ECG methods, at the level of 90 to 99%, with few works reporting an accuracy rate of 70% [16]. Many of them are based on public sets, and even in the case of using one public dataset, the accuracy of the methods between different studies may vary between 70 and 99% [16]. This may be the result of the richness of the data used in the studies. Also, it is important to remember that due to the complex spatial-temporal data structure, EEG research can often suffer from non-reproducibility due to inadequate data analysis methods and overfitting [94].

HRV focuses on changes in the beat-to-beat interval, which are ANS activity-dependent [95,96], while ECG signal morphology should be understood as the shape of the voltage curve over time. The morphology of the ECG represents different phases of depolarization and repolarization of the heart's sinuses and ventricles. The authors in [97] analyze together different psychiatric disorders, including schizophrenia, and present various degrees of classification accuracy from 73.7 to 89.2%, based on the selected ECG lead, and accuracy of 96.3% for 12 combined ECG leads. The fact that different levels of accuracy are achieved for different ECG leads may suggest that the classifier takes into account something different from brain activity. An additional limitation of this approach lies in the analysis of ECG morphology, as it increases the likelihood of classifying patients based on the side effects and physiological impacts of administered psychiatric medications. Different classes of drugs prescribed for specific psychiatric disorders exert distinct effects on various aspects of ECG morphology [98]. This limitation is less pronounced when HRV or R-R intervals are analyzed, as this type of signal lacks information about electrical potentials or specific time-dependent events, although it is also susceptible to medication-induced changes [75]. Another issue comes from the fact that the ECG morphology of different leads depends on the shape of the thoracic chest [99]. In that context, schizophrenic patients are a distinct group when it comes to their metabolic functioning and, thus, the presented body type [100]. For these reasons, we consider classification based solely on ECG morphology to be inadequate for the psychiatric diagnosis. In addition, a complex 12-lead ECG signal requires more complex equipment than a single-lead Polar H10.

The above-mentioned example shows mainly that the acquisition of complex signals may not always be a desired solution to the given problem, particularly in a situation where we do not have direct access to the reasoning behind the given conclusion, as in the case of DL algorithms [86]. This could also be a problem for EEG whole brain resting state signals.

HRV and ERP address the described problem by isolating and extracting distinct functional activities from the base signal, reducing its complexity. Thus, these approaches may reduce the risk of inadequate data analysis and overfitting [94]. In the case of HRV, this extraction appears to be without disadvantages as the rest of the ECG signal outside of the R-R intervals is not an established significant information carrier of brain activity. This lack of disadvantages does not apply to ERP as it is extracted from an EEG data source that represents brain activity as a whole, resulting in a loss of potentially significant data.

Finally, when evaluating the EEG methodology in relation to ECG in psychiatry, it is crucial to note that both resting-state EEG and ERP reflect cortical activity [85], with the ERP focusing on smaller-scale networks, such as the temporal–prefrontal network in the case of MMN [101]. In contrast, changes in HRV measured by the ECG represent the ANS activity regulated by deep brain structures [95,102] and provide information on the affective states mediated by ANS activity [96]. Taking into account this information, ECG and EEG analysis do not overlap in their utility and could potentially be treated as complementary methods that provide different types of insight into brain activity.

4.6. Wearables in the context of other disease monitoring approaches

In comparison to other diagnostic and treatment monitoring methods in healthcare, wearable devices tend to be the most cost-effective. This depends on the context of the disease we choose to monitor and the general availability of systemic healthcare [103]. In the case of ECG measurements with the use of Polar H10, at a price point lower than $100, this tool offers a significantly more affordable option compared to other advanced diagnostic methods. In addition, it is less time-consuming and readily available with high reusability, without the need for highly specialized personnel for it to be operational. With the proper software, it is ready to use at home.

As wearables mostly rely on general physiological signals, they are not self-sufficient in their diagnostic and monitoring tasks, especially in psychiatric healthcare. They are a source of transdiagnostic information [104] and require contextual interpretation through software or a trained physician. But the same applies even to other more robust diagnostic methods, e.g. neuroimaging (MRI) [105].

A very important disease monitoring approach in medicine in general is laboratory blood testing, including genetic profiling. In the case of psychiatry, this method presents challenges in interpretation due to the absence of established diagnostic biomarkers and the high variability in genetically determined traits, which complicates the integration of genetic tests into routine clinical practice [25].

This raises the question of the area of application of these methods. In contrast to the above-mentioned, wearable HRV monitoring is possibly much more publicly available. This allows it to be used outside of a clinical setting, for example, in rural areas and low-income countries, where the availability of professional healthcare is lacking as a first-line method of selective screening [106]. Another strategy would be the wide use of the method for population screening for psychotic disorders that could enable a faster diagnosis of the disease in the prodromal stage, reducing the burden on systemic healthcare. As shown in [46], good quality monitoring can be achieved even in a non-standardized at-home setting, potentially broadening the application beyond healthcare facilities (e.g. for patients with limited mobility, living in a remote location, or under-home quarantine).

Another aspect is the monitoring of disease treatment and progression using telemedicine, as HRV is correlated with the severity of psychotic diseases [39]. This would enable the execution of treatment plans in more controlled ways as one of the main treatment problems is low compliance and the recurrence of symptoms due to the patient losing continuity of treatment and contact with professional healthcare [107]. It would also potentially allow for intervention before the aggravation of symptoms.

Such large-scale proceedings would not be possible using neuroimaging or EEG, which are available primarily in specialized hospital settings, and not even laboratory testing.

4.7. Conclusions

In this study, we presented a classification approach, using short-duration R-R interval windows based on ECG signals, to differentiate between schizophrenia/bipolar disorder patients and healthy control individuals.

Our primary contribution lies in evaluating the accuracy of a classification-driven wearable-based diagnostic protocol. We examined the publicly available dataset containing samples from 60 persons. We implemented a state-of-the-art classification framework with a rigorous two-level cross-validation scheme for unbiased patient evaluation and automatic hyperparameter optimization. In addition, we compared several cutting-edge classification techniques, including deep learning approaches, to provide a comprehensive assessment.

Our method-independent approach, which yields an accuracy of 83% in 5-fold cross-validation and 80% in leave-one-out cross-validation, represents a significant step toward real-world clinical application. Our study indicates that wearable devices may be a cost-effective diagnostic tool, both for in- and outpatients. These promising results suggest that our framework could support future developments in diagnostic tools that integrate traditional symptom-based criteria with physiological markers, e.g. derived from ECG signals, for more precise and reliable diagnoses. In addition, our experiments correspond to a real-world scenario, assuming an almost unsupervised environment.

Furthermore, biomarker-based diagnostics supported by machine learning offers a potential pathway to improve the predictive validity and early interventions for schizophrenia and bipolar disorder in their prodromal phase. Our study opens the door to self-monitoring of the patient, even when performing daily activities, when the ECG is collected over a longer time period. In addition to early detection, this may provide an opportunity to monitor the development of the disease.

Author contributions

Conceptualization: Kamil Książek, Wilhelm Masarczyk, Przemysław Głomb, Magdalena Piegza.

Data curation: Kamil Książek.

Formal analysis: Kamil Książek, Michał Romaszewski, Krisztián Buza, Przemysław Sekuła, Michał Cholewa.

Funding acquisition: Kamil Książek.

Investigation: Kamil Książek, Michał Romaszewski, Krisztián Buza, Przemysław Sekuła, Michał Cholewa.

Methodology: Kamil Książek, Wilhelm Masarczyk, Przemysław Głomb, Michał Romaszewski, Krisztián Buza, Przemysław Sekuła, Michał Cholewa.

Project administration: Piotr Gorczyca, Magdalena Piegza.

Resources: Wilhelm Masarczyk, Piotr Gorczyca, Magdalena Piegza.

Software: Kamil Książek, Michał Romaszewski, Krisztián Buza, Przemysław Sekuła, Michał Cholewa.

Supervision: Magdalena Piegza.

Validation: Kamil Książek, Wilhelm Masarczyk, Michał Romaszewski, Krisztián Buza, Przemysław Sekuła, Michał Cholewa.

Visualization: Kamil Książek, Michał Romaszewski.

Writing – original draft: Kamil Książek, Wilhelm Masarczyk, Przemysław Głomb, Michał Romaszewski, Krisztián Buza, Przemysław Sekuła, Michał Cholewa, Katarzyna Kołodziej.

Writing – review & editing: Kamil Książek, Wilhelm Masarczyk, Przemysław Głomb, Michał Romaszewski, Krisztián Buza, Przemysław Sekuła, Michał Cholewa, Katarzyna Kołodziej.

References

1. Jain A, Mitra P. Bipolar Disorder. StatPearls Publishing. 2023. http://www.ncbi.nlm.nih.gov/pubmed/36579354
2. Solmi M, Seitidis G, Mavridis D, Correll CU, Dragioti E, Guimond S, et al. Incidence, prevalence, and global burden of schizophrenia - data, with critical appraisal, from the Global Burden of Disease (GBD) 2019. Mol Psychiatry. 2023;28(12):5319–27. https://doi.org/10.1038/s41380-023-02138-4 PMID: 37500825
3. Evans S, Banerjee S, Leese M, Huxley P. The impact of mental illness on quality of life: a comparison of severe mental illness, common mental disorder and healthy population samples. Qual Life Res. 2006;16(1):17–29. https://doi.org/10.1007/s11136-006-9002-6
4. Yamada Y, Matsumoto M, Iijima K, Sumiyoshi T. Specificity and continuity of schizophrenia and bipolar disorder: relation to biomarkers. Curr Pharm Des. 2020;26(2):191–200. https://doi.org/10.2174/1381612825666191216153508 PMID: 31840595
5. Larson MK, Walker EF, Compton MT. Early signs, diagnosis and therapeutics of the prodromal phase of schizophrenia and related psychotic disorders. Expert Rev Neurother. 2010;10(8):1347–59. https://doi.org/10.1586/ern.10.93 PMID: 20662758
6. Naser AY, Dahmash EZ, Alqahtani JS, Alsairafi ZK, Alsaleh FM, Alwafi H. Trends in hospital admissions for mental, behavioural and neurodevelopmental disorders in England and Wales between 1999 and 2019: an Ecological Study. Healthcare (Basel). 2022;10(11):2191. https://doi.org/10.3390/healthcare10112191 PMID: 36360532

7. Bessaha ML, Shumway M, Smith ME, Bright CL, Unick GJ. Predictors of hospital length and cost of stay in a national sample of adult patients with psychotic disorders. Psychiatr Serv. 2017;68(6):559–65. https://doi.org/10.1176/appi.ps.201600312 PMID: 28142382

8. Anderson KK, Norman R, MacDougall A, Edwards J, Palaniyappan L, Lau C, et al. Effectiveness of early psychosis intervention: comparison of service users and nonusers in population-based health administrative data. Am J Psychiatry. 2018;175(5):443–52. https://doi.org/10.1176/appi.ajp.2017.17050480 PMID: 29495897

9. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. American Psychiatric Association Publishing; 2022. https://doi.org/10.1176/appi.books.9780890425787

10. World Health Association. International Classification of Diseases 11th Revision (ICD-11). World Health Organization. 2018.

11. Pies R. How “objective” are psychiatric diagnoses?: (guess again). Psychiatry (Edgmont). 2007;4(10):18–22. PMID: 20428307

12. Strauss J. Subjectivity and severe psychiatric disorders. Schizophr Bull. 2011;37(1):8–13. https://doi.org/10.1093/schbul/sbq116 PMID: 20961994

13. Baca-Garcia E, Perez-Rodriguez MM, Basurte-Villamor I, Fernandez del Moral AL, Jimenez-Arriero MA, Gonzalez de Rivera JL, et al. Diagnostic stability of psychiatric disorders in clinical practice. Br J Psychiatry. 2007;190:210–6. https://doi.org/10.1192/bjp.bp.106.024026 PMID: 17329740

14. Howes OD, Cummings C, Chapman GE, Shatalina E. Neuroimaging in schizophrenia: an overview of findings and their implications for synaptic changes. Neuropsychopharmacology. 2023;48(1):151–67. https://doi.org/10.1038/s41386-022-01426-x PMID: 36056106

15. Hibar DP, Westlye LT, Doan NT, Jahanshad N, Cheung JW, Ching CRK, et al. Cortical abnormalities in bipolar disorder: an MRI analysis of 6503 individuals from the ENIGMA Bipolar Disorder Working Group. Mol Psychiatry. 2018;23(4):932–42. https://doi.org/10.1038/mp.2017.73 PMID: 28461699

16. Rahul J, Sharma D, Sharma LD, Nanda U, Sarkar AK. A systematic review of EEG based automated schizophrenia classification through machine learning and deep learning. Front Hum Neurosci. 2024;18:1347082. https://doi.org/10.3389/fnhum.2024.1347082 PMID: 38419961

17. Rawat K, Sharma T. PsyneuroNet architecture for multi-class prediction of neurological disorders. Biomedical Signal Processing and Control. 2025;100:107080. https://doi.org/10.1016/j.bspc.2024.107080

18. Wu Y, Su Y-A, Zhu L, Li J, Si T. Advances in functional MRI research in bipolar disorder: from the perspective of mood states. Gen Psychiatr. 2024;37(1):e101398. https://doi.org/10.1136/gpsych-2023-101398 PMID: 38292862

19. Wang C, Ren Y, Zhang R, Wang C, Ran X, Shen J, et al. Schizophrenia classification and abnormalities reveal of brain region functional connection by deep-learning multiple sparsely connected network. Biomedical Signal Processing and Control. 2024;96:106580. https://doi.org/10.1016/j.bspc.2024.106580

20. Owen MJ, Legge SE, Rees E, Walters JTR, O’Donovan MC. Genomic findings in schizophrenia and their implications. Mol Psychiatry. 2023;28(9):3638–47. https://doi.org/10.1038/s41380-023-02293-8 PMID: 37853064

21. Stogios N, Gdanski A, Gerretsen P, Chintoh AF, Graff-Guerrero A, Rajji TK. Autonomic nervous system dysfunction in schizophrenia: impact on cognitive and metabolic health. 2021.

22. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. Nat Rev Genet. 2006;7(2):130–41. https://doi.org/10.1038/nrg1769 PMID: 16418748

23. Bullmore E, Barnes A, Bassett DS, Fornito A, Kitzbichler M, Meunier D, et al. Generic aspects of complexity in brain imaging data and other biological systems. Neuroimage. 2009;47(3):1125–34. https://doi.org/10.1016/j.neuroimage.2009.05.032 PMID: 19460447

24. Smith SM, Nichols TE. Statistical challenges in “Big Data” human neuroimaging. Neuron. 2018;97(2):263–8. https://doi.org/10.1016/j.neuron.2017.12.018 PMID: 29346749

25. Torrey EF. Did the human genome project affect research on Schizophrenia?. Psychiatry Res. 2024;333:115691. https://doi.org/10.1016/j.psychres.2023.115691 PMID: 38219345

26. Dabiri M, Dehghani Firouzabadi F, Yang K, Barker PB, Lee RR, Yousem DM. Neuroimaging in schizophrenia: a review article. Front Neurosci. 2022;16:1042814. https://doi.org/10.3389/fnins.2022.1042814 PMID: 36458043

27. World Health Organization. Mental disorders. 2022. https://www.who.int/news-room/fact-sheets/detail/mental-disorders

28. WHO T. Mental health atlas 2020. World Health Organization. 2021. https://www.who.int/publications/i/item/9789240036703

29. Wee ZY, Yong SWL, Chew QH, Guan C, Lee TS, Sim K. Actigraphy studies and clinical and biobehavioural correlates in schizophrenia: a systematic review. J Neural Transm (Vienna). 2019;126(5):531–58. https://doi.org/10.1007/s00702-019-01993-2 PMID: 30888511

30. Markiewicz R, Markiewicz-Gospodarek A, Dobrowolska B. Galvanic skin response features in psychiatry and mental disorders: a narrative review. Int J Environ Res Public Health. 2022;19(20):13428. https://doi.org/10.3390/ijerph192013428 PMID: 36294009

31. Pelegrino A, Guimaraes AL, Sena W, Emele N, Scoriels L, Panizzutti R. Dysregulated noradrenergic response is associated with symptom severity in individuals with schizophrenia. Front Psychiatry. 2023;14:1190329. https://doi.org/10.3389/fpsyt.2023.1190329 PMID: 38025452

32. Benjamin BR, Valstad M, Elvsåshagen T, Jönsson EG, Moberget T, Winterton A, et al. Heart rate variability is associated with disease severity in psychosis spectrum disorders. Prog Neuropsychopharmacol Biol Psychiatry. 2021;111:110108. https://doi.org/10.1016/j.pnpbp.2020.110108 PMID: 32946948

33. Cella M, Okruszek Ł, Lawrence M, Zarlenga V, He Z, Wykes T. Using wearable technology to detect the autonomic signature of illness severity in schizophrenia. Schizophr Res. 2018;195:537–42. https://doi.org/10.1016/j.schres.2017.09.028 PMID: 28986005

34. Hickey BA, Chalmers T, Newton P, Lin C-T, Sibbritt D, McLachlan CS, et al. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: a systematic review. Sensors (Basel). 2021;21(10):3461. https://doi.org/10.3390/s21103461 PMID: 34065620

35. Welch V, Wy TJ, Ligezka A, Hassett LC, Croarkin PE, Athreya AP, et al. Use of mobile and wearable artificial intelligence in child and adolescent psychiatry: scoping review. J Med Internet Res. 2022;24(3):e33560. https://doi.org/10.2196/33560 PMID: 35285812

36. Hinde K, White G, Armstrong N. Wearable devices suitable for monitoring twenty four hour heart rate variability in military populations. Sensors (Basel). 2021;21(4):1061. https://doi.org/10.3390/s21041061 PMID: 33557190

37. Lieberman JA, First MB. Psychotic disorders. N Engl J Med. 2018;379(3):270–80. https://doi.org/10.1056/NEJMra1801490 PMID: 30021088

38. Pearlson GD. Etiologic, phenomenologic, and endophenotypic overlap of schizophrenia and bipolar disorder. Annu Rev Clin Psychol. 2015;11:251–81. https://doi.org/10.1146/annurev-clinpsy-032814-112915 PMID: 25581236

39. Książek K, Masarczyk W, Głomb P, Romaszewski M, Stokłosa I, Ścisło P, et al. Assessment of symptom severity in psychotic disorder patients based on heart rate variability and accelerometer mobility data. Comput Biol Med. 2024;176:108544. https://doi.org/10.1016/j.compbiomed.2024.108544 PMID: 38723395

40. Buza K, Książek K, Masarczyk W, Głomb P, Gorczyca P, Piegza M. A simple and effective classifier for the detection of psychotic disorders based on heart rate variability time series. In: ITAT'23: Information Technologies - Applications and Theory, 2023. p. 217–22.

41. Reinertsen E, Osipov M, Liu C, Kane JM, Petrides G, Clifford GD. Continuous assessment of schizophrenia using heart rate and accelerometer data. Physiol Meas. 2017;38(7):1456–71. https://doi.org/10.1088/1361-6579/aa724d PMID: 28653659

42. Inoue T, Shinba T, Itokawa M, Sun G, Nishikawa M, Miyashita M, et al. The development and clinical application of a novel schizophrenia screening system using yoga-induced autonomic nervous system responses. Front Physiol. 2022;13:902979. https://doi.org/10.3389/fphys.2022.902979 PMID: 36277195

43. Schlier B, Krkovic K, Clamor A, Lincoln TM. Autonomic arousal during psychosis spectrum experiences: Results from a high resolution ambulatory assessment study over the course of symptom on- and offset. Schizophr Res. 2019;212:163–70. https://doi.org/10.1016/j.schres.2019.07.046 PMID: 31422861

44. Johar S, Manjula GR. Interfering sensed input classification model using assimilated whale optimization and deep Q-learning for remote patient monitoring. Biomedical Signal Processing and Control. 2024;93:106202. https://doi.org/10.1016/j.bspc.2024.106202

45. Cella M, He Z, Killikelly C, Okruszek Ł, Lewis S, Wykes T. Blending active and passive digital technology methods to improve symptom monitoring in early psychosis. Early Interv Psychiatry. 2019;13(5):1271–5. https://doi.org/10.1111/eip.12796 PMID: 30821079

46. Tomasi J, Zai CC, Zai G, Herbert D, Richter MA, Mohiuddin AG, et al. Investigating the association of anxiety disorders with heart rate variability measured using a wearable device. J Affect Disord. 2024;351:569–78. https://doi.org/10.1016/j.jad.2024.01.137 PMID: 38272363

47. Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability. Circulation. 1996;93(5):1043–65. https://doi.org/10.1161/01.CIR.93.5.1043

48. Stracina T, Ronzhina M, Redina R, Novakova M. Golden standard or obsolete method? Review of ECG applications in clinical and experimental context. Front Physiol. 2022;13:867033. https://doi.org/10.3389/fphys.2022.867033 PMID: 35547589

49. Fernandes BS, Williams LM, Steiner J, Leboyer M, Carvalho AF, Berk M. The new field of "precision psychiatry". BMC Med. 2017;15(1):80. https://doi.org/10.1186/s12916-017-0849-x PMID: 28403846

50. Soares-Weiser K, Maayan N, Bergman H, Davenport C, Kirkham AJ, Grabowski S, et al. First rank symptoms for schizophrenia (cochrane diagnostic test accuracy review). Schizophr Bull. 2015;41(4):792–4. https://doi.org/10.1093/schbul/sbv061 PMID: 25939661

51. Książek K, Masarczyk W, Głomb P, Romaszewski M, Stokłosa I, Ścisło P, et al. HRV-ACC: a dataset with R-R intervals and accelerometer data for the diagnosis of psychotic disorders using a Polar H10 wearable sensor. 2023. https://doi.org/10.5281/zenodo.8171266

52. Rzewuska M. Validity and reliability of the Polish version of the Positive and Negative Syndrome Scale (PANSS). Int J Methods Psychiatr Res. 2002;11(1):27–32. https://doi.org/10.1002/mpr.120 PMID: 12459802

53. Zhao B, Lu H, Chen S, Liu J, Wu D. Convolutional neural networks for time series classification. Journal of Systems Engineering and Electronics. 2017;28(1):162–9.

54. Kourou K, Rigas G, Papaloukas C, Mitsis M, Fotiadis DI. Cancer classification from time series microarray data through regulatory Dynamic Bayesian Networks. Comput Biol Med. 2020;116:103577. https://doi.org/10.1016/j.compbiomed.2019.103577 PMID: 32001012

55. Mor B, Garhwal S, Kumar A. A systematic review of hidden Markov models and their applications. Arch Computat Methods Eng. 2020;28(3):1429–48. https://doi.org/10.1007/s11831-020-09422-4

56. Buza K. ASTERICS: projection-based classification of EEG with asymmetric loss linear regression and genetic algorithm. In: 2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI). 2020. p. 35–40. https://doi.org/10.1109/saci49304.2020.9118837

57. Pisner DA, Schnyer DM. Support vector machine. In: Mechelli A, Vieira S, editors. Machine Learning. Academic Press; 2020. p. 101–21.

58. Jankowski D, Jackowski K, Cyganek B. Learning decision trees from data streams with concept drift. Procedia Computer Science. 2016;80:1682–91. https://doi.org/10.1016/j.procs.2016.05.508

59. Tomašev N, Buza K, Marussy K, Kis PB. Hubness-aware classification, instance selection and feature construction: survey and extensions to time-series. Studies in Computational Intelligence. Berlin, Heidelberg: Springer; 2014. p. 231–62. https://doi.org/10.1007/978-3-662-45620-0_11

60. Xi X, Keogh E, Shelton C, Wei L, Ratanamahatana CA. Fast time series classification using numerosity reduction. In: Proceedings of the 23rd International Conference on Machine Learning - ICML '06. 2006. p. 1033–40. https://doi.org/10.1145/1143844.1143974

61. Keogh E, Wei L, Xi X, Vlachos M, Lee S-H, Protopapas P. Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures. The VLDB Journal. 2008;18(3):611–30. https://doi.org/10.1007/s00778-008-0111-4

62. Wang X, Mueen A, Ding H, Trajcevski G, Scheuermann P, Keogh E. Experimental comparison of representation methods and distance measures for time series data. Data Min Knowl Disc. 2012;26(2):275–309. https://doi.org/10.1007/s10618-012-0250-5

63. Goodfellow I, Bengio Y, Courville A. Deep Learning. MIT Press. 2016.

64. Christ M, Braun N, Neuffer J, Kempa-Liehr AW. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). Neurocomputing. 2018;307:72–7. https://doi.org/10.1016/j.neucom.2018.03.067

65. Overview on extracted tsfresh features. Tsfresh documentation. 2025. https://tsfresh.readthedocs.io/en/latest/text/list_of_features.html

66. Chapelle O, Zien A. Semi-supervised classification by low density separation. In: International Workshop on Artificial Intelligence and Statistics. 2005. p. 57–64.

67. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016. p. 785–94. https://arxiv.org/abs/1603.02754

68. Elsayed N, S A, Bayoumi M. Deep gated recurrent and convolutional network hybrid model for univariate time series classification. IJACSA. 2019;10(5). https://doi.org/10.14569/ijacsa.2019.0100582

69. Karim F, Majumdar S, Darabi H, Chen S. LSTM fully convolutional networks for time series classification. IEEE Access. 2018;6:1662–9. https://doi.org/10.1109/access.2017.2779939

70. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems. 2017;30.

71. Sorella S, Lapomarda G, Messina I, Frederickson JJ, Siugzdaite R, Job R, et al. Testing the expanded continuum hypothesis of schizophrenia and bipolar disorder. Neural and psychological evidence for shared and distinct mechanisms. Neuroimage Clin. 2019;23:101854. https://doi.org/10.1016/j.nicl.2019.101854 PMID: 31121524

72. Dines M, Kes M, Ailán D, Cetkovich-Bakmas M, Born C, Grunze H. Bipolar disorders and schizophrenia: discrete disorders?. Front Psychiatry. 2024;15:1352250. https://doi.org/10.3389/fpsyt.2024.1352250 PMID: 38745778

73. Faurholt-Jepsen M, Kessing LV, Munkholm K. Heart rate variability in bipolar disorder: a systematic review and meta-analysis. Neurosci Biobehav Rev. 2017;73:68–80. https://doi.org/10.1016/j.neubiorev.2016.12.007 PMID: 27986468

74. Valkonen-Korhonen M, Tarvainen MP, Ranta-Aho P, Karjalainen PA, Partanen J, Karhu J, et al. Heart rate variability in acute psychosis. Psychophysiology. 2003;40(5):716–26. https://doi.org/10.1111/1469-8986.00072 PMID: 14696725

75. Hattori S, Kishida I, Suda A, Miyauchi M, Shiraishi Y, Fujibayashi M, et al. Effects of four atypical antipsychotics on autonomic nervous system activity in schizophrenia. Schizophr Res. 2018;193:134–8. https://doi.org/10.1016/j.schres.2017.07.004 PMID: 28709776

76. Smith R, Thayer JF, Khalsa SS, Lane RD. The hierarchical basis of neurovisceral integration. Neurosci Biobehav Rev. 2017;75:274–96. https://doi.org/10.1016/j.neubiorev.2017.02.003 PMID: 28188890

77. Berntson GG, Bigger JT Jr, Eckberg DL, Grossman P, Kaufmann PG, Malik M, et al. Heart rate variability: origins, methods, and interpretive caveats. Psychophysiology. 1997;34(6):623–48. https://doi.org/10.1111/j.1469-8986.1997.tb02140.x PMID: 9401419

78. Appelhans BM, Luecken LJ. Heart rate variability as an index of regulated emotional responding. Review of General Psychology. 2006;10(3):229–40. https://doi.org/10.1037/1089-2680.10.3.229

79. Pumprla J, Howorka K, Groves D, Chester M, Nolan J. Functional assessment of heart rate variability: physiological basis and practical applications. Int J Cardiol. 2002;84(1):1–14. https://doi.org/10.1016/s0167-5273(02)00057-8 PMID: 12104056

80. Henry BL, Minassian A, Paulus MP, Geyer MA, Perry W. Heart rate variability in bipolar mania and schizophrenia. J Psychiatr Res. 2010;44(3):168–76. https://doi.org/10.1016/j.jpsychires.2009.07.011 PMID: 19700172

81. Clamor A, Lincoln TM, Thayer JF, Koenig J. Resting vagal activity in schizophrenia: meta-analysis of heart rate variability as a potential endophenotype. Br J Psychiatry. 2016;208(1):9–16. https://doi.org/10.1192/bjp.bp.114.160762 PMID: 26729841

82. Ortiz A, Bradler K, Moorti P, MacLean S, Ishrat Husain M, Sanches M, et al. Increased sympathetic tone is associated with illness burden in bipolar disorder. J Affect Disord. 2022;297:471–6. https://doi.org/10.1016/j.jad.2021.10.089 PMID: 34715156

83. Barà C, Pernice R, Catania CA, Hilal M, Porta A, Humeau-Heurtier A, et al. Comparison of entropy rate measures for the evaluation of time series complexity: simulations and application to heart rate and respiratory variability. Biocybernetics and Biomedical Engineering. 2024;44(2):380–92.

84. Cura OK, Akan A, Atli SK. Detection of attention deficit hyperactivity disorder based on EEG feature maps and deep learning. Biocybernetics and Biomedical Engineering. 2024;44(3):450–60.

85. Newson JJ, Thiagarajan TC. EEG frequency bands in psychiatric disorders: a review of resting state studies. Front Hum Neurosci. 2019;12:521. https://doi.org/10.3389/fnhum.2018.00521 PMID: 30687041

86. Ranjan R, Sahana BC, Bhandari AK. Deep learning models for diagnosis of Schizophrenia using EEG signals: emerging trends, challenges, and prospects. Arch Computat Methods Eng. 2024;31(4):2345–84. https://doi.org/10.1007/s11831-023-10047-6

87. Ruiz de Miras J, Ibáñez-Molina AJ, Soriano MF, Iglesias-Parro S. Schizophrenia classification using machine learning on resting state EEG signal. Biomedical Signal Processing and Control. 2023;79:104233. https://doi.org/10.1016/j.bspc.2022.104233

88. Silva CP da, Tedesco S, O'Flynn B. EEG datasets for healthcare: a scoping review. IEEE Access. 2024;12:39186–203. https://doi.org/10.1109/access.2024.3376254

89. Hamilton HK, Boos AK, Mathalon DH. Electroencephalography and event-related potential biomarkers in individuals at clinical high risk for psychosis. Biol Psychiatry. 2020;88(4):294–303. https://doi.org/10.1016/j.biopsych.2020.04.002 PMID: 32507388

90. Tada M, Kirihara K, Mizutani S, Uka T, Kunii N, Koshiyama D, et al. Mismatch negativity (MMN) as a tool for translational investigations into early psychosis: a review. Int J Psychophysiol. 2019;145:5–14. https://doi.org/10.1016/j.ijpsycho.2019.02.009 PMID: 30831138

91. Erickson MA, Ruffle A, Gold JM. A meta-analysis of mismatch negativity in Schizophrenia: from clinical risk to disease specificity and progression. Biol Psychiatry. 2016;79(12):980–7. https://doi.org/10.1016/j.biopsych.2015.08.025 PMID: 26444073

92. Raggi A, Serretti A, Ferri R. The P300 component of the auditory event-related potential in adult psychiatric and neurologic disorders: a narrative review of clinical and experimental evidence. Int Clin Psychopharmacol. 2025;40(5):259–74. https://doi.org/10.1097/YIC.0000000000000566 PMID: 39163164

93. Campanella S. Use of cognitive event-related potentials in the management of psychiatric disorders: Towards an individual follow-up and multi-component clinical approach. World J Psychiatry. 2021;11(5):153–68. https://doi.org/10.5498/wjp.v11.i5.153 PMID: 34046312

94. Kinahan S, Saidi P, Daliri A, Liss J, Berisha V. Achieving Reproducibility in EEG-Based Machine Learning. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency. 2024. p. 1464–74. https://doi.org/10.1145/3630106.3658983

95. ten Donkelaar HJ, Němcová V, Lammens M, Overeem S. The autonomic nervous system. Clinical neuroanatomy. Springer; 2020. p. 669–710. https://doi.org/10.1007/978-3-030-41878-6_12

96. Gullett N, Zajkowska Z, Walsh A, Harper R, Mondelli V. Heart rate variability (HRV) as a way to understand associations between the autonomic nervous system (ANS) and affective states: a critical review of the literature. Int J Psychophysiol. 2023;192:35–42. https://doi.org/10.1016/j.ijpsycho.2023.08.001 PMID: 37543289

97. Tasci B, Tasci G, Dogan S, Tuncer T. A novel ternary pattern-based automatic psychiatric disorders classification using ECG signals. Cogn Neurodyn. 2024;18(1):95–108. https://doi.org/10.1007/s11571-022-09918-8 PMID: 38406197

98. O'Brien P, Oyebode F. Psychotropic medication and the heart. Adv psychiatr treat. 2003;9(6):414–23. https://doi.org/10.1192/apt.9.6.414

99. Kania M, Rix H, Fereniec M, Zavala-Fernandez H, Janusek D, Mroczka T, et al. The effect of precordial lead displacement on ECG morphology. Med Biol Eng Comput. 2014;52(2):109–19. https://doi.org/10.1007/s11517-013-1115-9 PMID: 24142562

100. Sugawara N, Yasui-Furukori N, Tsuchimine S, Fujii A, Sato Y, Saito M, et al. Body composition in patients with schizophrenia: comparison with healthy controls. Ann Gen Psychiatry. 2012;11(1):11. https://doi.org/10.1186/1744-859X-11-11 PMID: 22554352

101. Garrido MI, Kilner JM, Stephan KE, Friston KJ. The mismatch negativity: a review of underlying mechanisms. Clin Neurophysiol. 2009;120(3):453–63. https://doi.org/10.1016/j.clinph.2008.11.029 PMID: 19181570

102. Tiwari R, Kumar R, Malik S, Raj T, Kumar P. Analysis of heart rate variability and implication of different factors on heart rate variability. Curr Cardiol Rev. 2021;17(5):e160721189770. https://doi.org/10.2174/1573403X16999201231203854 PMID: 33390146

103. De Sario Velasquez GD, Borna S, Maniaci MJ, Coffey JD, Haider CR, Demaerschalk BM, et al. Economic perspective of the use of wearables in health care: a systematic review. Mayo Clin Proc Digit Health. 2024;2(3):299–317. https://doi.org/10.1016/j.mcpdig.2024.05.003 PMID: 40206120

104. Beauchaine TP, Thayer JF. Heart rate variability as a transdiagnostic biomarker of psychopathology. Int J Psychophysiol. 2015;98(2 Pt 2):338–50. https://doi.org/10.1016/j.ijpsycho.2015.08.004 PMID: 26272488

105. Alexandros Lalousis P, Wood S, Reniers R, Schmaal L, Azam H, Mazziota A, et al. Transdiagnostic structural neuroimaging features in depression and psychosis: a systematic review. Neuroimage Clin. 2023;38:103388. https://doi.org/10.1016/j.nicl.2023.103388 PMID: 37031636

106. Saxena S, Thornicroft G, Knapp M, Whiteford H. Resources for mental health: scarcity, inequity, and inefficiency. Lancet. 2007;370(9590):878–89. https://doi.org/10.1016/S0140-6736(07)61239-2 PMID: 17804062

107. Biringer E, Hartveit M, Sundfør B, Ruud T, Borg M. Continuity of care as experienced by mental health service users - a qualitative study. BMC Health Serv Res. 2017;17(1):763. https://doi.org/10.1186/s12913-017-2719-9 PMID: 29162112